波尔图大学工程学院

基于视频的音乐生成

Serkan Sulun

arXiv:2602.07063v1 [cs.LG] 5 Feb 2026



电气与计算机工程博士项目

导师：Paula Viana 博士

导师：Matthew E. P. Davies 博士

2025年5月23日

© Serkan Sulun，2025年

基于视频的音乐生成

Serkan Sulun

电气与计算机工程博士项目

经委员会口试审查通过：

主席：João Paulo Trigueiros da SilvaCunha 博士波尔图大学工程学院委员：Rui Pedro Pinto de Carvalho ePaiva 博士科英布拉大学科学与技术学院委员：Tom Collins 博士迈阿密大学弗罗斯特音乐学院委员：Sofia Carmen Faria Maia Cavaco 博士里斯本新大学科学与技术学院委员：Maria Teresa Magalhães da SilvaPintode Andrade 博士波尔图大学工程学院委员（导师）：Paula Maria Marques de MouraGomes Viana 博士*ISEP*，波尔图理工学院

2025年5月23日

摘要

随着互联网上视频内容的数量迅速增长，寻找合适的配乐仍是一项重大挑战。本论文介绍了 EMSYNC（情感与同步），这是一种快速、免费且自动化的解决方案，可根据输入视频生成量身定制的音乐，使内容创作者无需作曲或获取音乐授权即可提升作品品质，从而简化创作与制作流程。我们的模型能够生成与视频在情感和节奏上同步的音乐，为自动配乐生成提供了一种自适应且富有表现力的解决方案。

EMSYNC 的核心组成部分之一是一种新颖的视频情绪分类器。为了实现准确且高效的视频分类，我们巧妙地融合了预训练模型。此外，我们通过使用大型数据集开展电影预告片类型分类实验，解决视频分类中以数据为中心的挑战。通过利用预训练深度神经网络进行特征提取，并在训练期间冻结这些网络、仅训练融合层，我们降低了计算复杂度，同时提高了准确率。我们的方法在 Ekman-6 和 MovieNet 上取得了最先进的结果，展现出良好的泛化能力；这两个数据集分别是用于情感分类和电影流派分类的最大视频数据集。

另一项重要贡献是构建了一个大规模、带有情感标注的 MIDI 数据集，用于情感音乐生成。我们利用在线资源中的标注，建立了规模最大的带有 效价-唤醒度标签的 MIDI 数据集。此外，我们还分析了 MIDI 文件中歌曲歌词的情绪内容。随后，我们介绍了一种基于情感的 MIDI 生成器，它是首个以连续情感值而非离散类别作为条件的生成器，从而能够生成细腻且契合复杂情绪内容的音乐。

为增强时间同步，我们提出了一种新颖的时间边界条件控制方法，称为“边界偏移编码”，用于使音乐和弦与场景变化保持同步。该方法集成到 EMSYNC 中，确保音乐自然地跟随视频的节奏和韵律，从而提升整体用户体验。

我们还探讨了音频合成，鉴于成对 MIDI-音频数据稀缺，重点关注音频带宽增强。我们提出了一项概念验证，以揭示并应对音频合成中的挑战，尤其强调泛化能力。我们首次指出了“滤波器过拟合”问题：在特定低通滤波器上训练的模型无法泛化到真实场景。为解决这一问题，我们提出了一种数据增强策略，其效果优于标准正则化方法，为开发适用于真实场景的稳健音频增强模型迈出了第一步。

结合视频情感分类、基于情绪的音乐生成和时间边界条件化，EMSYNC 成为一款全自动的基于视频的音乐生成器。用户研究表明，在音乐丰富度、情感对齐、时间同步和总体偏好方面，它始终优于现有方法。因此，EMSYNC 为基于视频的音乐生成树立了新的最先进水平，创作出的音乐在情感和节奏上都与视频相契合。

摘要

互联网上的视频内容呈指数级增长，使大量且极具价值的资源向大众开放。然而，对于这些内容的创作者而言，找到合适的配乐仍是一项不小的挑战。本论文介绍了EMSYNC（情感与同步），这是一种快速、免费且自动化的解决方案，可根据输入视频生成自适应音乐，让内容创作者无需作曲或购买音乐授权，就能提升作品质量，从而促进创意与制作。我们的模型能够生成在情感和节奏上与视频同步的音乐，为自动配乐生成提供了一种自适应且富有表现力的方案。EMSYNC 的核心组件之一是一种新的视频情感分类器。为实现准确、高效的分类，我们提出了一种智能的预训练模型融合方法。为克服适用于该任务的数据稀缺问题，并考虑到目标相似性，我们利用了与电影预告片类型分类任务相关的大型数据集。通过使用预训练深度神经网络提取特征，并在训练过程中冻结这些网络、仅训练融合层，我们降低了计算复杂度并提高了准确率。我们在 Ekman-6 和 MovieNet 数据集上验证了该方法的泛化能力，结果超越了当前最先进水平；这两个数据集分别是用于视频情感分类和电影类型分类的最大视频数据集。

本论文的另一项重要贡献是一个带情感标注的MIDI数据集。通过结合从在线资源收集的信息以及对歌曲歌词情感内容的分析，我们构建了规模最大的、带有效价—唤醒度标注的MIDI数据集，从而为训练基于情感的音乐生成模型提供了强大的资源。论文的另一项核心贡献是一个能够根据情感生成MIDI的模型，它首次以连续情感值而非离散类别为条件。这种方法能够生成与复杂情感内容相契合的音乐。为改善音频与视频内容之间的时间同步，我们提出了一种新的时间边界条件方法，称为“边界偏移编码”，可使音乐和弦与视频中的场景变化相匹配。该方法集成于 EMSYNC 中，确保音乐自然地跟随视频的节奏和速度，从而提升用户体验。

本论文还探讨了音频合成。鉴于多乐器MIDI-音频数据配对稀缺，我们聚焦于与带宽增强相关的方面。我们提出了一项概念验证，旨在识别并应对音频合成中的主要挑战，尤其关注泛化能力。我们发现了“滤波器过拟合”问题：使用特定低通滤波器训练的模型无法很好地泛化到真实场景。为解决这一问题，我们提出了一种数据增强策略，其效果优于标准正则化方法，这标志着

pri开发可用于实际场景的鲁棒音频增强模型的第一步 .

通过结合视频情感分类、基于情感的音乐生成和时间条件控制，EMSYNC 成为一款基于视频的自动音乐生成器。用户研究表明，在音乐丰富度、情感对齐、时间同步和总体偏好方面，它始终优于现有方法。因此，EMSYNC 在基于视频的音乐生成领域树立了新的最先进水平，创作出在情感和节奏上与视频相契合的音乐。

致谢

首先，我谨向我的导师 Paula Viana 和 Matthew Davies 致以诚挚的谢意，感谢他们一直以来的支持与指导。我也永远感激我的家人——Ali、Lale 和 Zeynep——给予我无条件的爱与支持。他们为我打下了坚实的基础，使我能够建立独立的生活，而不必在成年生活常有的艰辛之上，还背负童年创伤的重担。

人们常说，读博士期间很难维系感情，但有一位了不起的女朋友，让这段时光对我而言轻松了许多。我永远也无法充分感谢我的爱人、伴侣和最好的朋友 Chloé。在这个异国他乡，她为我带来了一个新的家。

特别感谢 Bianca，她不仅是一位了不起的朋友，还为我介绍了教学工作，更重要的是，她包容了我的种种抱怨。我也要感谢在波尔图的好友兼学术同仁 Göksu、Nathalie 和 Nabila，感谢他们始终如一的支持、友谊和宝贵建议。此外，也感谢你们所有人接纳我加入“波尔图团队”。

我非常幸运，拥有好友 Doˇgan。尽管我们身处不同国家，过去 20 年来仍一直保持着亲密的友谊。我很感激同事兼办公室伙伴 Inês、Luís、Eduardo、Américo 和 Tiago 的友谊与支持，他们帮助我度过了在葡萄牙的时光。

最后，我想向 la Caixa奖学金同届的博士同学——Gizem、Marta、Inês、Arturo、Javi、Chris、Jacob、Paula、Jaime 和 Ludovica——表达谢意，感谢他们的友谊，以及一起组织了几次令人难忘的假期。

我也很荣幸获得了两项奖学金，分别是 la Caixa博士奖学金（INPhINIT）（LCF/BQ/DI19/11730032）和 FCT（科学与技术基金会）博士奖学金（2022.09594.BD），使我得以开展这项博士研究。此外，NEXUS-2项目（12407/BI-M-ED_B2/2025）也提供了资金支持。

Serkan Sulun

目录

1 引言 1

1.1 符号音乐生成.31.1.1 时间条件音乐生成.41.1.2 基于情绪的音乐生成.51.2 视频分析.61.3 基于视频的音乐生成.91.4 音频生成与带宽扩展.101.5 贡献.11

2 背景与相关工作 15

2.1 符号音乐.162.2 情绪表征.182.2.1 类别式情感表示.182.2.2 维度式情感表示.192.3 Transformer.20

2.4 序列处理.252.4.1 文本情感分类.272.4.1.1 数据集.272.4.1.2 方法.282.4.2 条件通用序列生成.292.4.3 条件符号音乐生成.302.4.3.1 数据集.302.4.3.2 方法.322.5 视频分析.332.5.1 视频情感分类.342.5.1.1 数据集.342.5.1.2 方法.342.5.2 预告片类型分类.352.5.2.1 数据集.352.5.2.2 方法.362.6 基于视频的符号音乐生成.372.7 音频带宽扩展.392.7.1 数据集.402.7.2 方法.40

2.8 后续章节概述.42

3 MIDI 情绪标注 4 **3**
3.1 使用 Spotify 特征.443.2 歌曲歌词的情绪分类.483.2.1 训练.483.2.1.1 数据集.
483.2.1.2 实现细节.493.2.2 推理.49

3.2.3 结果.50
3.2.3.1 在 GoEmotions 数据集上的情绪分类 . . . . . 503.2.3.2 在 MIDI 数据集上的推理.523

.3 讨论. 56

4 条件音乐生成 **57**
4.1 基于时间边界的音乐生成.574.2 基于情绪的音乐生成.604.3 数据集与预处理.
624.4 实现细节.634.5 评估.644.6 结果与讨论.66

5 视频分析 **68**
5.1 场景切换提取.685.2 语义视频分类.695.2.1 特征提取.695.2.2 情感分类.725.2.2.1 数据集与特征提取.725.2.2.2 分类方法.735.2.2.3 实现细节与超参数.735.2.2.4 数据集清洗.745.2.2.5 推理Web应用.755.2.2.6 结果与讨论.755.2.3 流派分类.775.2.3.1 数据集与特征提取.775.2.3.2 分类方法.785.2.3.3 实现细节与超参数.825.2.3.4 结果与讨论.84

6 基于视频的音乐生成 88
6.1 情感映射.886.2 实施细节.896.3 评估.936.3.1 数据集.946.3.2 客观评估.946.3.3 主观评估.956.4 结果与讨论.95

7 音频带宽扩展 10 **1**
7.1 模型.103
7.1.1 U-Net.1037.1.2 ResNet.1037.2 正则化方法.1057.2.1 随机失活.1057.2.2 批归一化.1057.2.3 数据增强.1067.3 数据集.106

7.4 评估.1087.4.1 指标.1087.4.2 测试.109

7.4.3 验证.1107.5 实现细节.1107.6 结果.1107.6.1 验证数据.1107.6.2 测试数据.1127.6.3 样本过拟合.1137.6.4 U-Net 与 ResNet：模型比较.1147.6.5 带宽扩展可视化.1147.7 讨论.116

8 结论 **118**
8.1 贡献.118
8.2 未来方向.121

参考文献 **124**

图目录

1.1 我们的基于视频的音乐生成方法.3 2.1 乐谱（左）、MIDI（中）和钢琴卷帘（右）表示。摘自 Müller（2015年）.16 2.2 钢琴卷帘到基于事件的表示的转换。摘自 Oore 等人（2020年）。许可：CC 4.0 BY.17 2.3 Plutchik（1988年）的情绪轮。作者/版权持有者：Machine Elf 1735。版权条款和许可：公有领域.19 2.4 情绪环形模型（Russell，1980）.20 2.5 Transformer 模型，左侧为编码器，右侧为解码器。-https://github.com/dvgodoy/dl-visuals/?tab=readme-ov-file，CC BY 4.0 . . . . 21 2.6 正弦位置编码。Cosmia Nebula 创作 - 自创作品，CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=119948133 . . . . . 23 3.1 数据集创建流程.45 3.2 我们的 7 标签（左）和 28 标签（右）数据集中各情绪对应的样本数量.53 4.1 音乐边界及其偏移量的图示.58 4.2 基于边界偏移量的条件机制.59 4.3 基于情感的音乐生成模型.62 4.4 情感表征评估流程.66 5.1 视频情绪分类流程.74 5.2 在 Ekman-6 数据集测试集上，按真实标签归一化数值的混淆矩阵.76 5.3 用于流派分类的视频特征提取流程.77 5.4 我们的多层感知机模型.78 5.5 我们的单变换器模型.79 5.6 我们的多变换器模型.81 5.7 我们的基线模型。下标“I”和“A”分别指图像和音频.82 5.8 仅使用 CLIP 特征时，帧数与平均精度均值的关系。. . . . . 87 6.1 我们的基于视频的音乐生成流程。箭头旁的数值为样本值.90 6.2 边界图示.91 6.3 我们的音乐生成器。效价和唤醒度下方的数字为样本值。 92 6.4 我们的基于视频的音乐生成调查问卷截图.96

6.5 具有自我报告音乐理论理解的参与者的堆叠条形图结果.976.6 不具有自我报告音乐理论理解的参与者的堆叠条形图结果.986.7 所有参与者的堆叠条形图结果.987.1 所用的音频带宽扩展模型。（左）U-Net 模型，虚线表示堆叠连接。（右）包含 15 个残差块的 ResNet 模型。c、k 和 s 分别表示卷积层的通道大小、卷积核大小和步幅.1047.2 训练滤波器的频率响应。每个图中均叠加了未见滤波器（6 阶巴特沃斯滤波器）的频率响应.1087.3 模型在训练过程中的验证性能.1117.4 音频片段样本的频谱图.1157.5 与目标频谱图的绝对差值。为提高可见度，对色彩图进行了反转.116

表格目录

2.1 文本情感分类数据集.282.2 MIDI 数据集.312.3 视频情感数据集.352.4 全频带音频数据集.41

3.1 带情感标签的 MIDI 数据集.463.2 GoEmotions 标签与 Ekman类别的映射.493.3 7标签分类结果.513.4 28标签分类结果.513.5 7标签数据集中的样本条目.533.6 28标签数据集中的样本条目.544.1 我们的方法在评估期间的演奏表现.664.2 我们的方法在推理期间的演奏表现.67

5.1 与当前最先进水平在 Ekman-6 数据集上的分类准确率。765.2 我们的模型与当前最先进水平的演奏表现对比.845.3 各类型的结果及宏平均值。MN：MovieNet（Q. Huang 等人，2020年），BL：基线（我们的方法），MT：多变换器（我们的方法）.855.4 纳入预训练特征对平均精确率和运行时间的影响。86

6.1 Ekman（1971年）类别情绪对应的效价和唤醒度均值与标准差，取自 Russell 和 Mehrabian（1977）的用户研究.896.2 具有自我报告的音乐理论理解的参与者的结果。较低的值表示更好的排名（1 = 最佳， 3 = 最差）。值以均值（标准差）表示.966.3 不具有自我报告的音乐理论理解的参与者的结果。较低的值表示更好的排名（1 = 最佳， 3 =最差）。值以均值（标准差）表示.966.4 所有参与者的结果。较低的值表示更好的排名（1 = 最佳， 3 =最差）。值以均值（标准差）表示.977.1 两种不同训练设置下所使用的低通滤波器类型和顺序： 单滤波器 （无数据增强）和多滤波器 （数据增强）。1077.2 测试数据集上的输出信噪比（SNR）和绝对 VGG 距离（VGG），以及相对于输入的改进.113

7.3 我们的基线模型在不采用任何正则化方法时，分别在训练数据划分和测试数据划分上的输出信噪比（SNR），以及与输入信噪比的比较.1147.4 我们模型的参数数量、乘加运算次数（MACs）和运行时间.114

缩略语列表

AAC 情感算法作曲 AEC 声学回声消除 应用程序接口 应用程序编程接口 自动语音识别 音频频谱图 Transformer BERT 来自 Transformer 的双向编码器表示 BEATs 音频变换器双向编码器表征 批归一化 批归一化 每分钟节拍数 每分钟节拍数 CLIP 对比式语言-图像预训练 CMT 可控音乐 Transformer 卷积神经网络 卷积神经网络 CTRL 条件 Transformer 语言 数据增强 数据增强 数字音频工作站 数字音频工作站 Demixing Secrets 数据集 Demixing Secrets数据集 深度神经网络 随机失活 随机失活 EMotion and SYNChronization EMotion and SYNChronization EMOPIA 情绪钢琴 Fréchet 音频距离 全连接层 全连接 GPT 生成式预训练 门控循环单元 门控循环单元 I3D 双流膨胀三维卷积网络 ICASSP 声学、语音与信号处理国际会议 大语言模型 大语言模型 Lakh MIDI 数据集 Lakh MIDI 数据集 线性预测编码 Lakh 钢琴卷帘数据集 Lakh 钢琴卷帘数据集 长短期记忆网络 长短期记忆网络

乘累加 平均精度均值 平均精度均值 MAESTRO 用于同步音轨与组织的MIDI和音频 乐器数字接口 乐器数字接口 多层感知机 百万歌曲数据集 百万歌曲数据集 音乐视频情感数据集 音乐视频情感数据集 非数值 非数值 负对数似然 负对数似然 自然语言处理 自然语言处理 光学字符识别 光学字符识别 精确率 精确率 成对情感关系识别 PERR 语音质量感知评估 召回率 召回率 ResNet 残差网络 循环神经网络 循环神经网络 信噪比 信噪比 超分辨率 时间关系网络 时间关系网络 时间段网络 VEMOCLAP 使用预训练特征的视频情绪分类器 VGMIDI 电子游戏乐器数字接口 视觉几何组 视觉几何组 虚拟语音质量客观听评器 视觉 Transformer 视频视觉Transformer YOLO You Only Look Once

第1章

引言

用户生成多媒体内容的指数级增长，源于价格实惠的高质量录音设备和视频编辑软件日益普及（Falkowski-Gilski 与 Uhl，2020年）。内容创作中的一项重大挑战是选择合适的配乐以提升观众参与度（Buhler，2018年）。然而，未经授权使用已商业出版的音乐会侵犯版权，从而限制在 YouTube 等平台上的变现。购买音乐、聘请作曲家或搜索免版税音轨等替代方案可能成本高昂、耗时，或无法确保与视频的适当同步。本论文提出一种快速、免费且免版税的解决方案：一种模型，以用户提供的任意视频作为输入，并自动创作与视频情绪内容相匹配的配乐。

音乐可以用两种主要格式表示：音频和符号。音频是普通听众所感知的标准音乐格式，将声音表示为波形。符号音乐则主要由音乐家使用，乐谱便是其典型形式，它编码音高、时值，并包含渐强或极弱等演奏指示。在数字领域，符号音乐被表示为类似于文本的离散序列。使用最广泛的数字符号音乐格式是 MIDI（乐器数字接口）。MIDI文件由一系列事件组成，每个事件编码有关音符的信息，例如其音高、乐器（通道）以及强度（力度/响度）。虽然音乐作品可以像乐谱一样存储为 MIDI 文件，但 MIDI 文件必须合成为音频才能被听到。Fluidsynth[1] 等软件使用标准库为 MIDI 事件分配声音，并将其转换为音频波形。

现有以音频波形形式生成音乐的方法缺乏可编辑性（Copet 等人，2023年；Evans 等人，2024；Lam 等人，2023年；H. Liu 等人，2024）。这些方法将音乐制作的所有阶段——作曲、演奏、编辑、混音和母带处理——压缩为一个单一流程，限制了专业人士的创作掌控力。相比之下，以 MIDI 等符号格式生成音乐则更加灵活。由于 MIDI 会明确编码音符、时序和力度等音乐元素，专业人士可以轻松润饰音乐作品、使用数字音频工作站（DAW）合成音频，或录制并调整演奏。

[1] https://fluidsynth.com

我们的工作聚焦于使用深度神经网络（DNN）从任意视频生成 MIDI 格式的音乐。由于 MIDI 是我们唯一使用的格式，因此我们交替使用“MIDI”和“音乐”这两个词。训练视频到MIDI的深度神经网络面临的一项关键挑战，是缺少大规模配对视频-MIDI数据集。现有数据集要么仅针对特定领域，例如钢琴家手部视频（Koepke 等人，2020年），要么仅包含数量有限的样本（约 1k）（Di等人，2021；Zhuo 等人，2023年）。虽然存在 Lakh MIDI 数据集等大型 MIDI 数据集，但它们没有与视频配对，因此无法进行端到端学习（Raffel，2016）。我们通过确定连接视频与音乐的中间表示来解决这一问题，由此构建了一个两阶段框架：首先，分析模型从输入视频中提取表征；然后，这些表征用于对音乐生成模型进行条件设定。

为了实现一款能够吸引人类听众的自动配乐生成器，我们必须模拟人类配乐作曲家的作曲过程。因此，中间表示应与作曲家创作配乐的方式相一致。如何创作适合视频的音乐是一个由来已久的问题（Buhler，2018年）。然而，这与其说是工程问题，不如说是艺术问题——它没有客观或明确的答案。尽管如此，某些启发式方法和普遍认知表明，有效的配乐应当：（1）与视频蕴含的情感相契合，以增强这些情感（Deutsch，2007；Schifrin，2011年）；（2）匹配视频的时间动态，确保同步，从而提升观众参与度（Cohen，1990；Prendergast，1992）。基于这一观点，我们的方法采用双分支方法，利用情感线索和时间动态生成音乐。图1.1 展示了我们的双阶段、双分支方法。

首先分析用户提供的输入视频，然后根据分析结果生成符号音乐，最后生成与符号音乐相对应的音频。然而，在我们的研究以及本论文的结构安排中，我们对这三个过程采用了更具策略性的顺序。

在我们的研究和本论文的叙述中，我们都从音乐生成开始，随后进行视频分析。这样的顺序可确保生成音乐在每个阶段都保持感知上的连贯性和音乐上的意义。如果先设计视频分析模块，风险会很大：如果音乐生成器无法有效利用提取出的表征，我们就需要重新审视并重新设计视频分析模块。此外，一旦建立了条件音乐生成器，即使视频特征提取未能成功，我们也可以采用备用方案，直接从用户那里获取这些特征。

设计完音乐生成和视频分析模块后，我们着重将二者整合起来。对于 MIDI 转音频生成，我们采用标准方案，使用专门的 Fluidsynth 软件和标准音色库。虽然已有基于深度神经网络的实验方法，但这些方法仅适用于钢琴音轨，无法推广至多乐器音乐（Hawthorne 等人，2019年；Sambaragi 等人，2024）。

不过，在本论文接近结尾处，我们会在不同的上下文中探索深度神经网络用于音频生成的可能性。这项探索独立于我们的主要模型，并未整合其中。我们专门聚焦于音频生成所面临的挑战。正如后文所述，

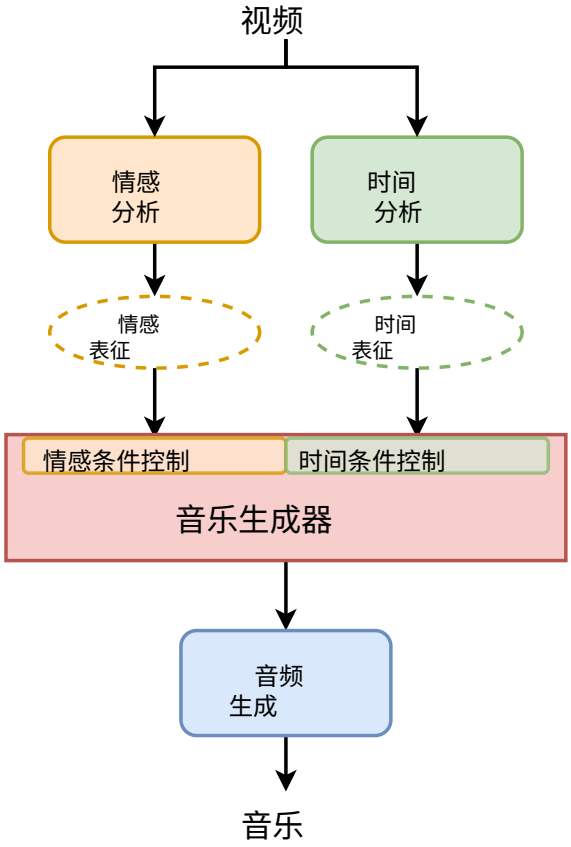


图 1.1：我们的基于视频的音乐生成方法。

第 1.4节，由于缺少包含配对音频和多乐器 MIDI 的数据集，我们只能依赖纯音频数据集，尤其是在从低质量来源生成高质量音频时。在此上下文中，我们着手解决音频带宽扩展问题：通过对高质量音频进行合成带限处理来创建输入-目标对。

总而言之，以下是我们在本文中探讨的主题的大致顺序：

1. 符号音乐生成

2. 视频分析 3. 结合视频分析与音乐生成模块（基于视频的音乐生成）4. 音频生成与带宽扩展

我们首先讨论符号音乐生成。

1.1 符号音乐生成

符号音乐以音符序列的形式表示，因其紧凑且结构化的格式而广泛应用于机器学习模型。大规模原始 MIDI 数据集支持无监督

训练深度神经网络以实现自动符号音乐生成（Raffel，2016）。目前最先进的序列生成架构是 Transformer 模型，它是所有大语言模型（LLM）的骨干网络（Vaswani 等人，2017年）。与语言建模类似，Transformer 会学习预测下一个词元，即下一个音符，并在推理过程中每次生成一个词元，以自回归方式生成输出。然而，人类作曲家并非只是按顺序一个接一个地创作音符；他们的创作过程涉及动机、主题、情绪以及节奏等时间结构等高层次概念（Meyer，2008）。受此启发，我们提出一种条件音乐生成器，采用两种不同机制，分别融入时间条件和情感条件。

1.1.1 时间条件音乐生成

我们首先探讨基于时间动态生成音乐这一挑战。然而，首先有必要界定“时间动态”的含义。以往关于基于视频的音乐生成的研究，使用运动速度和运动显著性作为时间特征（Di等人，2021）。这种方法对时间动态进行密集表征，因此会持续对音乐音符密度进行条件化。然而，我们对Di等人（2021）研究开展的初步听辨测试揭示了两个关键局限。首先，不断变化的音符密度会破坏节奏一致性。其次，由于映射是密集的，除非相邻片段之间的运动速度和音符密度存在强烈对比，否则很难感知音乐与视频之间的同步。

我们提出以下研究问题：

> 在保持节奏一致性的同时，哪种类型的时间特征最有助于实现视频与音乐的同步？

我们的假设如下：

> 密集的时间表征（如速度）可能导致节奏不断发生变化和中断。

> 匹配稀疏时间特征（如时间边界），有助于在增强视频与音乐之间感知同步性的同时，保持节奏稳定。

MIDI生成模型中的时间条件控制面临着独特的挑战。深度Transformer模型可以处理视频等基于时间的数据，其序列维度会因固定帧率而与时间呈线性相关（Arnab 等人，2021；Z. Liu 等人，2022）。相比之下，MIDI 通常采用基于事件的编码进行表征，其中序列维度与时间维度并非线性相关（Oore 等人，2020年）。

在基于事件的 MIDI 编码中，主要有两种词元类型：“音符”和“时间偏移”。音符词元表示音高，而时间偏移词元则推进时间轴，从而记录音符时值和静音。每个时间偏移词元都会指定一个时间增量，形成一维

序列，其中序列中的位置并不直接对应时间上的位置。这种编码的主要优势在于没有固定的时间网格，因此能够灵活表现细微且富有表现力的时序变化，体现人类音乐家的演奏特质。

我们进一步提出：

> 我们如何设计不依赖词元位置、同时保留基于事件的 *MIDI* 编码的时间同步方法？

我们提出以下假设：

> 以时间条件形式提供的辅助输入可以促进同步。

当要求模型在特定时间戳进行同步时，会出现一个特殊挑战，因为强制同步可能会破坏生成音乐的一致节奏。为了解决这一问题，我们进一步研究了仅需最少干预的学习型条件化方法，使模型能够选择与其生成的音乐和谐一致的确切同步点。此外，我们假设，向模型提供有关未来边界的信息，可以引导模型以连贯的方式生成出现在边界之前的音符。

1.1.2 基于情绪的音乐生成

音乐长期以来一直是情感表达与交流的有力媒介（Hunter & Schellenberg，2010）。音乐唤起情绪的能力已在多个学科领域得到研究，包括心理学（Krumhansl，2002年）、音乐学（Juslin & Sloboda，2001）和神经科学（Koelsch，2014年）。随着深度学习的发展，人们对能够自动分析和生成音乐、以引发听众特定情绪反应的机器学习算法越来越感兴趣（Briot 等人，2020年）。

我们研究如何从情绪生成音乐，旨在尽可能提高性能。我们采用深度学习的发展趋势，即使用更多数据训练更大的模型，以获得更好的表现（Kaplan 等人，2020年）。然而，目前包含情绪标签的 MIDI 数据集规模很小，样本数量不足 400 个（Ferreira & Whitehead，2019年；Hung 等人，2021；K. Zhao 等人，2019年）。这一局限给训练用于基于情绪的 MIDI 生成的大型高性能深度神经网络带来了重大挑战。

为解决数据稀缺问题，我们探讨以下研究问题：

> 如何获取由 *MIDI*-情感标签对构成的大规模数据集？
>
> 是否可以在无需人工输入的情况下，为现有 *MIDI* 数据集添加情绪标签？
>
> 其他模态（如音频或文本，例如歌曲歌词）是否包含与情绪相关的信息？
>
> 我们能否利用现有的音乐数据库和平台来识别情绪标签，并将其分配给相应的 *MIDI* 样本？

我们假设：

> 从其他相关模态中自动提取的情绪，尤其是从音频版本和内置歌曲歌词中提取的情绪，将与相应的 *MIDI* 样本相匹配。

由于这些模态之间存在固有差异，即音频与 MIDI、文本与 MIDI 之间的差异，这些标签将作为“弱”标签。然而，大量可用样本预计将弥补 MIDI 与情绪标签之间匹配较弱的问题。由于互联网上的大多数音乐内容以音频格式提供，我们旨在探索音乐平台，以识别与情绪相关的元数据。我们还利用语言模型来分析 MIDI 文件中歌词的情绪。

我们现在继续探讨视频分析。

1.2 视频分析

在视频分析中，表征通常根据其抽象程度和复杂性被分为高层或低层。低层视频特征直接从视频数据中提取，而不解释其内容。示例包括基于像素的特征，如 RGB 值和边缘信息（Canny，1986）、运动特征，如光流（Horn & Schunck，1981年）、纹理特征，如局部二值模式（Ojala 等人，1996），以及转场特征，如叠化和场景切换（Bieda 等人，2021）。

高层视频特征对视频内容提供了更抽象的解读，通常通过机器学习方法来捕捉。示例包括动作（Jhuang 等人，2013年）、物体（Redmon 等人，2016）、情绪（Soleymani 等人，2012年）和类型（Wehrmann & Barros，2017年）。

在这项工作中，我们同时利用低层级的场景切换和高层级语义信息。由于场景切换检测已是成熟的问题，我们直接使用 FFMpeg 软件[2]应用现有方法。我们的研究主要聚焦于语义视频分类，尤其是情绪分类，因为这对我们的情感驱动音乐生成器至关重要。由于我们旨在开发一款适用于任何类型视频的音乐生成器，因此特别关注任意视频的情绪分类，而非特定领域的视频，例如包含对话的视频（Gao 等人，2021）或有人脸的视频（Lee 等人，2019年）。

在语义视频分类中，与大多数视频分类任务一样，深度神经网络代表着当前最先进水平。这些模型主要将像素值和音频波形等原始数据作为输入。与依赖手工设计特征的旧方法不同，深度神经网络旨在通过其各层隐式提取高层特征。然而，这些模型通常包含数百万个参数，因此需要大量计算资源，包括时间、内存和大规模训练数据。

视频分类由于未压缩的视频帧序列规模庞大而面临额外挑战，这进一步增加了计算需求。为解决这一问题，先前的研究

[2] https://ffmpeg.org

仅处理固定且数量有限的帧或片段（Wehrmann & Barros，2017年；S. Zhao 等人，2020年）。然而，这种方法不可避免地会导致信息损失，从而可能损害分类准确率。随着更大模型的采用，处理原始视频的计算复杂度进一步增加——这一趋势多年来仍在持续扩展（Bernstein 等人，2021；M. Tan & Le，2019年）。当训练数据有限时，使用更大的输入维度和更大的模型也会增加过拟合的可能性（Defernez & Kemsley，1999）。

我们提出以下研究问题：

> 我们如何设计具有强大表征能力的高效视频分析模型，使其能够处理任意数量的输入帧，同时确保泛化能力？

我们的假设如下：

> 由于原始视频输入体量庞大，使用此类数据端到端训练模型的成本极高，而且很可能效率低下。
>
> 在各种多媒体任务上预训练的模型可以为视频分类提供紧凑且有意义的输入，从而提升演奏表现和效率。

使用预训练模型带来多重优势。首先，由于预训练模型以推理模式使用，其权重保持冻结，因此只需进行一次前向传播，而无需额外的反向传播，从而降低时间和内存复杂度。其次，它大幅减小了馈入后续正在训练的模型的输入规模。例如，一个非常小的视频帧包含 $224 \times 224 \times 3 = 163{,}968$ 个值，而最先进水平的图像分析模型 CLIP（对比语言-图像预训练）仅用 512 个值表示一幅图像（Radford 等人，2021）。随着视频时值的增加，这种规模缩减变得更为重要。第三，假设多媒体分析模型的预训练特征为视频分类任务提供了有价值且相关的信息，则使用它们便无需训练新模型以从原始数据中隐式提取类似特征。这减小了需要训练的后续模型的规模。最后，由于它减小了输入规模以及训练中的权重数量，也降低了过拟合的可能性（Defernez & Kemsley，1999）。

利用预训练网络的方法有多种，且彼此之间有很大重叠。这些方法通常称为迁移学习、微调或使用预训练特征（Niu 等人，2020年；C. Tan 等人，2018年）。这些术语的区别如下：迁移学习是指将一个在某项任务上训练的模型应用于另一项任务（Weiss 等人，2016）。微调是指针对新任务进一步训练迁移后的模型（Church 等人，2021）。在我们的研究中，我们使用预训练特征，即使用预训练模型对数据进行推理，然后将所得输出或激活值（特征）输入一个新模型进行训练（Reiss 等人，2021）。这些方法常用于视频

处理任务，例如人体活动识别（Ray & Kolekar，2024）、视频摘要（Khan 等人，2024）和视频推荐（A. Almeida 等人，2022年）。

我们将继续探讨以下研究问题：

端到端训练特定任务深度神经网络，与利用冻结的预训练模型相比，效果如何？

如何高效利用开源预训练网络，并将其组合用于视频分类？

我们的假设如下：

强大的多模态预训练模型可以在推理模式下使用，并将其输出与单独训练的轻量级神经网络融合，从而同时提升演奏效果和效率。

我们介绍了端到端训练的架构，以及融合预训练网络的模型，并比较了它们的性能。我们采用用于图像分析、光学字符识别、自动语音识别、音频事件分类、音乐分类和面部情绪分类的预训练模型，以纳入有关视觉场景、物体、文本、语音、音乐、音频和人脸的信息。这些预训练模型以推理模式使用，也就是说不会对其进行微调，因此显著降低了时间和内存需求。随后，我们使用训练好的神经网络融合预训练特征，并获得最终的分类结果。由于预训练模型能够从原始数据中提取有意义的特征，因此仅包含一层或两层的浅层融合网络就足以充当分类器，从而大幅降低计算复杂度、硬件需求和训练时间。

我们还从质量、数量及其对总体目标的适用性等方面，研究了各种语义视频分类数据集。我们探讨了能够提供最大多样性的视频数据集类型，以使我们的模型能够泛化到任意视频。我们认为，用户生成视频和电影视频是两个最佳候选。除法律方面的考虑外，从 YouTube[3] 和 Flickr[4]，等平台收集的用户生成视频对上传视频的类型几乎没有限制。同样，电影作为一种艺术形式，追求突破创意边界，因此包含广泛的视觉和叙事多样性。

我们的初步调查表明，缺乏带有情绪标签的高质量、大规模视频数据集。Ekman-6是最大的视频情绪分类基准之一，仅包含1,637个视频，并导致过拟合（B. Xu 等人，2018年）。使用少量且固定数量的输入帧可以缓解过拟合，但代价是信息损失（Wehrmann & Barros，2017年；S. Zhao 等人，2020年）。在本研究中，我们旨在解决数据稀缺与信息损失这些问题。我们提出以下研究问题：

还有哪些语义视频分类任务可以提供大量训练数据，以进一步缓解过拟合？

[3] https://youtube.com
[4] https://flickr.com

在数据充足的情况下，输入帧的数量与分类性能之间有什么关系？

我们进一步质疑使用数量少且固定的帧这一做法，并提出：

得益于 *IMDb* 等在线目录[5] 以及 *YouTube* 等视频托管平台（*Harper & Konstan*，*2016*；*Q.Huang* 等，*2020*年），电影视频数据集提供了大量带标签的视频样本。

通过使用紧凑的输入和大量数据，增加输入帧的数量将持续提升分类性能。

为此，我们探索了包含大量样本的电影数据集。借助其规模，我们可以解决过拟合问题、训练更大且更强大的模型，并使用更多输入帧，充分发挥提供给模型的信息的作用。

与视频情感分类相对应的电影领域任务是电影预告片类型分类。由于情感与类型分类都涉及离散类别，我们可以采用能够预测一组预定义标签上概率分布的分类器模型。这种相似性使我们能够使用共享的模型架构来处理这两项任务。

我们继续讨论如何整合符号音乐生成与视频分析模块，以实现基于视频的音乐生成器。

1.3 基于视频的音乐生成

最后，我们整合先前构建的各个组件——视频情绪分类器、FFMpeg 场景切换提取器、情绪条件机制和时间边界条件机制——以构建基于视频的音乐生成器。

视频情绪分类器与音乐生成器的情绪条件机制之间表征不匹配，这是一个关键挑战。我们的视频情绪分类器输出离散情绪的概率，而音乐生成器则使用连续的效价-唤醒度值来表征情绪（Russell，1980）。我们提出以下问题：

我们如何弥合类别型（离散）与维度（连续）情绪表征之间的差异？

我们的假设是：

心理用户研究的统计数据可能有助于开发这两种表征之间的映射方法。

为弥合这一差距，我们提出一种基于心理学研究（Russell & Mehrabian，1977）的简单映射方法。这些研究向参与者展示与离散情绪相关的情境

[5] https://imdb.com

，并请他们根据效价（愉快与不愉快）和唤醒度（强度）对感受到的情感进行评分。研究人员随后报告每种离散情绪类别的效价和唤醒度的均值与标准差。我们利用这些统计值构建高斯分布混合，其中每个分量对应一种离散情绪。然后，根据视频情绪分类器输出的概率分布对该混合分布加权，并从中采样得到最终的效价-唤醒度值。最后，我们将完整模型与现有的基于视频的音乐生成器进行比较评估。大量用户研究表明，在所有评估指标上，我们的方法均优于以往的方法。

虽然我们用户研究中的音频样本是使用现成的 Fluidsynth 软件生成的，但我们进一步探讨了音频生成任务，尤其关注带宽扩展及其相关挑战。

1.4 音频生成与带宽扩展

我们的视频音乐生成器生成的符号音乐紧凑、可编辑且易于理解。为了生成人类可以听到的音频，用户有多种选择，从快速但质量较低的方法，到高质量但耗时较长的方案。首先，用户可以使用 Fluidsynth 等计算机软件和音色库，几乎无需干预，便可快速将 MIDI 合成为音频。这些音色库为特定乐器的特定音高提供固定音色。由于我们的工作重点是音乐创作而非音色设计，因此我们采用这种方法生成最终样本。然而，正如预期的那样，与人类音乐家创作的音乐相比，生成的音频听起来机械、不真实，整体质量也较低。

另一种选择是使用数字音频工作站（DAW）将 MIDI 转换为音频。在这种方法中，最终用户可以编辑 MIDI 的各个方面，包括音符、速度、乐器和力度，也可以使用专有音色库、自定义音效以及各种混音和母带处理技术来编辑音频本身。虽然这种方法生成的音频质量更高，但需要手动操作，耗费大量时间、技能和劳动。最后，也可以由音乐家演奏并录制我们的音乐生成器生成的音乐作品。这种方法能获得最高质量的音频，但同样需要投入最多的时间、技能和劳动。

另一种面向研究的解决方案是使用深度神经网络将 MIDI 合成为音频。然而，这是一个定义不明确的问题，因为同一首音乐作品（即一个 MIDI文件）可以有无数种人类演奏方式。尽管有模型试图解决这一任务，但它们依赖于 MIDI 与音频配对的数据集，而这类数据集通常以钢琴为基础（Hawthorne 等人，2019年；Sambaragi 等人，2024）。由于我们的工作聚焦于多乐器音乐生成，我们面临的挑战是，目前没有包含人类演奏音频与 MIDI 配对数据的多乐器数据集。我们可以在包含 MIDI 及其对应合成音频的合成数据集上训练深度神经网络，但这会使深度神经网络的性能受限于

合成方法的质量，从而使其显得有些多余。此外，这也可以被视为一个泛化问题：模型只能生成听起来像合成音频的内容，无法泛化到音乐家创作的真实世界样本。

为了进一步研究这一问题，我们提出以下研究问题：

还有哪些音频生成任务会表现出由合成数据引起的过拟合？

我们如何量化并缓解音频生成任务中对合成数据的过拟合？

我们的假设如下：

音频增强任务也会因使用合成降质音频作为输入而出现过拟合。

为了系统地分析过拟合并减轻这一问题，我们可以在深度神经网络的训练和测试过程中使用多种不同的降质函数。

音频增强任务以低质量音频作为输入，并以对应的高质量音频作为输出。音频增强模型通常使用合成数据进行训练：利用数字滤波器对高质量音频进行降质，生成对应的低质量音频（Kuleshov 等，2017年）。这项任务只需要高质量音频数据集，无需来自其他模态（例如 MIDI）的对应样本。由于音频增强任务仅使用纯音频样本，因此数据十分充足。我们旨在通过应用多种数字降质滤波器，并衡量训练后的模型能否泛化到不同滤波器，来量化其泛化能力。具体而言，我们聚焦于音频带宽扩展任务；通过观察和评估缺失频带的重建质量，可以对其进行定性和定量分析。此外，在音频带宽扩展中，我们可以改变低通滤波器的类型和阶数，从而使用多种低通滤波器。

由于这个问题源于合成数据的使用，我们假设解决方案是采用数据增强策略，而非基于模型的方法。结果表明，在创建输入样本时使用多种滤波器可以缓解过拟合。我们的方法迈出了实现真实世界音频带宽扩展模型的第一步。

现在，我们总结一下我们的贡献。

1.5 贡献

我们的首项关键贡献是构建了一个带有情绪标签的大规模 MIDI 数据集。该数据集支持训练以情绪作为条件输入、以 MIDI 序列作为真实目标的模型。基于在更大数据集上训练更大模型能够取得更好结果这一广为认可的原则（Sevilla 等人，2022年），我们避免使用现有的情感标注 MIDI 数据集，因为其中的样本数量较少（Ferreira 和 Whitehead，2019年；Hung 等人，2021；K. Zhao 等人，2019年）。相反，我们为 Lakh MIDI 数据集分配情绪标签，该数据集

包含 176,581 个样本（Raffel，2016）。通过整合多个辅助数据集，我们为 34,791 个 MIDI 样本标注了效价和唤醒度值，创建了迄今规模最大的情感标注 MIDI 数据集。利用该数据集，我们成功训练了大型 Transformer 模型（Vaswani 等人，2017年），以生成经情绪条件化的多乐器符号音乐。据我们所知，这是首个能够同时基于效价和唤醒度进行条件化的音乐生成模型，从而可以控制广泛使用的情感环形模型中的任意情绪（Russell，1980）。我们在以下论文和开源代码中分享了研究成果。

> ***S. Sulun***、*M. E. P. Davies* 和 *P. Viana*，《基于连续值情感条件的符号音乐生成》，*IEEE Access*，第 *10* 卷，第 *44617–44626* 页，*2022*年。(Sulun 等人，2022年)
>
> 基于情感的 *MIDI* 生成器。https://github.com/serkansulun/midi-emotion

我们还通过对 MIDI 文件中包含的歌词进行情感分类，为 MIDI 文件添加情绪标签。为此，我们首先在 GoEmotions 上训练情绪分类模型；GoEmotions 是规模最大的文本数据集之一，包含 28 个细粒度情绪标签（Demszky 等人，2020年）。随后，我们利用这些模型从与 MIDI 文件相关的歌词中推断情绪标签。尽管这种方法未能提升基于情绪的 MIDI 生成效果，但我们的模型在 GoEmotions 测试集上达到了最先进的性能，且模型规模仅为基线模型的一半。我们在以下论文中介绍了研究发现，并公开了开源代码和数据集。

> ***S. Sulun***，*P. Oliveira* 与 *P. Viana*，《*Emotion4MIDI*：基于歌词的情绪标注符号音乐数据集》，载于 *Progress in Artificial Intelligence*，*Cham*：施普林格《自然》瑞士，*2023*年，第*77–89*页。（Sulun 等人，2023年）
>
> *MIDI*歌词的情绪分类。包含预训练情绪文本分类器，以及带有 *Lakh* 和 *Reddit MIDI*数据集中 *MIDI* 文件标签的数据集。
> https://github.com/serkansulun/lyricsemotions

我们还提出了一种用于音乐生成器时间条件控制的新机制。与依赖固定粗粒度时间网格的最先进水平基于视频的音乐生成器不同（Di等人，2021；Kang 等人，2024），我们的方法在保留基于事件的编码的同时实现了细粒度时间控制。我们通过编码并输入边界偏移量 到模型中来实现这一点。这些偏移量告知模型距离下一个和弦的剩余时间，使其能够“预判”即将出现的和弦，并相应地生成前置音符。因此，生成的和弦及周围音符在节奏与和声上保持连贯。我们在介绍总体模型的最终论文中提出了该方法，并附有在线演示。

> ***S. Sulun***、*P. Viana* 和 *M. E. P. Davies*，《通过对齐情绪与时间边界生成视频配乐》，*IEEE*多媒体汇刊，*2026*年，即将刊印。(Sulun等，2026)

基于视频的音乐生成器。https://github.com/serkansulun/emsync
演示。
https://colab.research.google.com/drive/1hq6WmZztJk-yKp0UX1UTon8rJ3xC-bV6

随后，我们介绍用于语义视频分类的方法。该方法利用紧凑的预训练特征而非原始帧，从而保持较低的计算复杂度。在视频情感分类中，我们使用交叉注意力层（Bahdanau 等人，2015年）对预训练特征进行智能融合。与最先进水平的模型相比，我们的方法在规模最大的情绪视频分类数据集之一 Ekman-6（B. Xu 等人，2018年）上取得了显著优势。作为一项次要贡献，我们发现 Ekman-6 数据集中存在标注错误。这些错误源于视频搜索过程中的问题，例如将名为“Joy”的名人视频误标为表达喜悦情绪。为解决这一问题，我们人工审查了 Ekman-6 数据集中的所有视频，并整理出一份错误标注样本黑名单，以供未来研究者参考。我们在以下论文和在线演示中展示了研究结果：

S. Sulun、*P. Viana* 和 *M. E. P.Davies*，《*VEMOCLAP*：视频情绪分类网页应用》，国际多媒体研讨会（*ISM*），*IEEE*计算机学会，第 *137–140* 页，*2024*。(Sulun 等人，2024年b)

视频情绪分类器。 https://github.com/serkansulun/video-emotion
演示。
https://colab.research.google.com/drive/1S-Nsm968-KTErbuU0qHOp7mEEzh6xOVx

虽然 Ekman-6 数据集包含 1,637 个视频，平均时长为 10 秒（B. Xu 等人，2018年），电影数据集 MovieNet 包含 33,000 部电影预告片，平均时长为 2 分钟（Q. Huang 等人，2020年）。为了使用 MovieNet 进行预告片类型分类，我们通过增加输入帧数量来训练更稳健的架构，同时避免过拟合。为了融合从视频帧和音频片段中提取的预训练特征，我们采用 Transformer 架构——它被公认为序列处理的最先进方法，也是现代大语言模型的基础（Vaswani 等人，2017年）。该架构非常适合处理长输入序列并捕捉长期依赖关系。借助这一能力，我们的类型分类模型会处理完整的视频和音频内容：首先识别场景切换，然后使用所有由此产生的帧以及完整音频。因此，我们的方法在 MovieNet（目前最大的电影数据集）上的流派分类表现优于现有最先进模型。此外，我们通过实证证明，随着纳入更多输入帧，分类性能会持续提升。我们将在下文介绍研究结果，并提供相应的开源代码。

S. Sulun、*P. Viana* 和 *M. E. P. Davies*，“使用多模态预训练特征进行电影预告片体裁分类”，《专家系统与应用》，第 *258* 卷，页码 *125209*，*2024* 年。(Sulun 等人，2024a)
预告片类型分类器。 https://github.com/serkansulun/trailer-genre-classification

我们最终提出了应用于音乐音频带宽扩展的深度神经网络的滤波器泛化问题。与深度学习所应用的许多问题不同，我们并未发现对音频样本本身（即训练数据）过拟合的任何证据，而是观察到最先进的深度神经网络明显倾向于对滤波器形状过拟合。当这些深度神经网络面对经过训练中未包含的低通滤波器预处理的音频片段时，无法获得有意义的带宽扩展。此外，使用批归一化和随机失活等广泛采用的正则化层也不足以缓解这一问题。为解决滤波器过拟合问题，我们提出一种新颖的数据增强方法，即在训练时使用多个滤波器。我们的结果表明，若不进行数据增强，滤波器过拟合会随着训练的推进而加剧，而引入数据增强则是实现滤波器泛化的一个有前景的步骤。我们的发现呈现在以下论文与开源代码中。

> ***S. Sulun***和 *M. E.P. Davies*，《关于使用深度神经网络进行音乐带宽扩展的滤波器泛化》，*IEEE* 信号处理选题期刊，第 *15* 卷，第 *1* 期，第 *132–142* 页，*2020*年。 (Sulun & Davies，2021)
> 音频带宽增强器。 https://github.com/serkansulun/deep-music-enhancer

本论文的其余部分结构如下。第2 章介绍情绪表征、Transformer、序列处理（文本和 MIDI）、视频分析、基于视频的音乐生成以及音频带宽扩展的背景与相关工作。第3 章详细介绍我们如何为大型 MIDI 数据集标注情绪。第4 章讨论我们的条件音乐生成方法，重点关注情感和时间边界条件控制。第5章介绍我们的视频分析模型，具体包括情绪分类、流派分类和场景切换提取。第6 章将这些组件整合为一个完整的基于视频的音乐生成模型。第7 章介绍我们对音频带宽扩展任务的探索。最后，第8 章总结我们的结论，并展望未来方向。

第 2 章

背景与相关工作

本章首先介绍与我们的基于视频的音乐生成目标相关的关键概念。我们的工作涉及多种类型的数据——视频、图像、音频、文本、符号音乐和情感。视频、图像、音频和文本具有众所周知且标准化的数字格式，而符号音乐和情感没有统一的表示方式。因此，我们首先介绍符号音乐和情感的表征，以及对它们进行编码的不同方式。

接下来，我们介绍 Transformer 架构，它是序列处理领域最先进的方法（Vaswani 等人，2017年）。由于符号音乐是由音符组成的序列，视频是由帧组成的序列，而文本是由单词或子词组成的序列，因此我们在各项子任务中广泛采用 Transformer。随后在论文中，我们将介绍对标准 Transformer 所作的重要改进，以实现新颖的条件机制和多模态处理。因此，首先从理论层面介绍 Transformer 的工作原理十分重要。

解释上述关键概念后，我们将回顾特定应用领域的相关文献。我们首先介绍序列处理这一通用任务，然后介绍其具体子任务，例如文本情感分类、条件通用序列生成和条件符号音乐生成。

接下来，我们将在单独的章节中介绍视频分析，将其与通用序列处理区分开来。尽管视频可以描述为由帧组成的序列，但其结构化的像素布局——涵盖宽度、高度和时间——使其有别于 MIDI 和文本等经过词元化的序列。本节还将讨论视频情感分类和预告片类型分类。

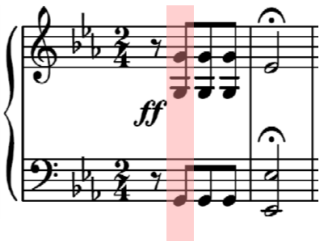
接下来，我们回顾基于视频的符号音乐生成相关文献，这代表了我们整体任务的当前最先进水平。最后，我们介绍音频带宽扩展方面的研究，它是音频生成的一个子任务。

2.1 符号音乐

符号音乐格式（例如 MIDI（乐器数字接口））用于在数字领域表示音乐演奏和音乐作品。这些文件仅包含音符、速度和力度等音乐信息，不包含实际声音，因此类似于“数字乐谱”。与音频格式相比，符号音乐文件的体积和维度都小得多，因此更易于管理，也更适合使用深度神经网络进行建模（Briot 等人，2020年）。目前最常用的两种音乐格式是 MIDI（乐器数字接口）和钢琴卷帘，两者可以无缝转换。

与音频格式（如 WAV、MP3）不同，MIDI 不存储实际声音，而是编码音高、力度、时值和乐器选择等音乐指令，这些指令可由兼容 MIDI 的设备（如合成器、虚拟乐器）进行解读。MIDI文件由事件组成，每个音符都以时间差值（表示时间偏移）、音高（表示音符）、通道（表示乐器）和力度（表示响度）进行编码。因此，它形成了一种一维序列表示。

相比之下，钢琴卷帘格式以二维表征组织音符：横轴表示时间，纵轴表示音高，值对应力度，由此构成一个二维矩阵。对于多乐器音乐，会使用多个二维矩阵，每种乐器对应一个。图 2.1 展示了贝多芬第五交响曲的前十二个音符，分别以乐谱、MIDI 和钢琴卷帘格式呈现。在此图中，乐谱和 MIDI 分别通过不同的谱表和通道表示钢琴的左手和右手音符。由于这些音符属于同一种乐器，钢琴卷帘将它们汇集到一个矩阵中。



Time (Ticks)	Message	Channel	Note Number	Velocity
60	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	67	100
0	NOTE ON	1	55	100
0	NOTE ON	2	43	100
55	NOTE OFF	1	67	0
0	NOTE OFF	1	55	0
0	NOTE OFF	2	43	0
5	NOTE ON	1	63	100
0	NOTE ON	2	51	100
0	NOTE ON	2	39	100
240	NOTE OFF	1	63	0
0	NOTE OFF	2	51	0
0	NOTE OFF	2	39	0

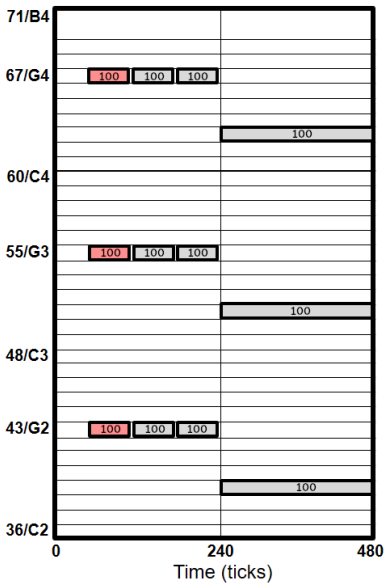


图 2.1：乐谱（左）、MIDI（中）和钢琴卷帘（右）的表征。摘自 Müller（2015年）。

基于事件的符号音乐表征是为神经网络应用而对 MIDI 所做的改编（Oore 等人，2020年）。在分词之前，它将 MIDI 的部分属性拆分为独立事件，从而创建一个规模适中的字典。具体而言，它将增量时间信息分离出来，形成时间偏移事件，并将力度信息分离出来，形成设置力度事件。时间偏移词元用于沿时间轴移动，表示音符时值以及音符之间的静音。每个词元都指定以毫秒为单位的时间增量。例如，800 毫秒的偏移编码为时间偏移<800毫秒>。力度和时间偏移都可以量化，以缩小词汇表大小。通常，时间偏移以 8 毫秒为单位量化，最大时间偏移为 1000 毫秒（Oore 等人，2020年）。超过 1000 毫秒的时间偏移可以用多个连续的时间偏移词元表示。

图 2.2 展示了将符号音乐转换为基于事件的表示的过程。如2.1所示，与 MIDI 相比，钢琴卷帘是一种更直观的表征，因此本图使用了钢琴卷帘。不过，同样的转换过程也适用于 MIDI。首先，演奏一个时值为 640 毫秒、力度为 31 的 C4音符，随后是 24 毫秒的静音。之后，演奏一个力度为 25 的 F3音符。在处理第一个音符时，假设之前的力度不同，我们使用 SET-VELOCITY<31> 词元将力度设为 31。然后，使用 NOTE-ON<C4> 词元标记第一个音符的开始及其力度。接下来，使用TIME-SHIFT<640ms>，将时间向前推进 640 毫秒，直到到达下一个音符边界（即 NOTE-ON 或NOTE-OFF 词元）。由于下一个音符的力度不同于当前设定的力度 31，我们使用 SET-VELOCITY<25>词元将新的力度设为 25。最后，使用 NOTE-ON<F3> 词元标记下一个音符的开始及其音高。

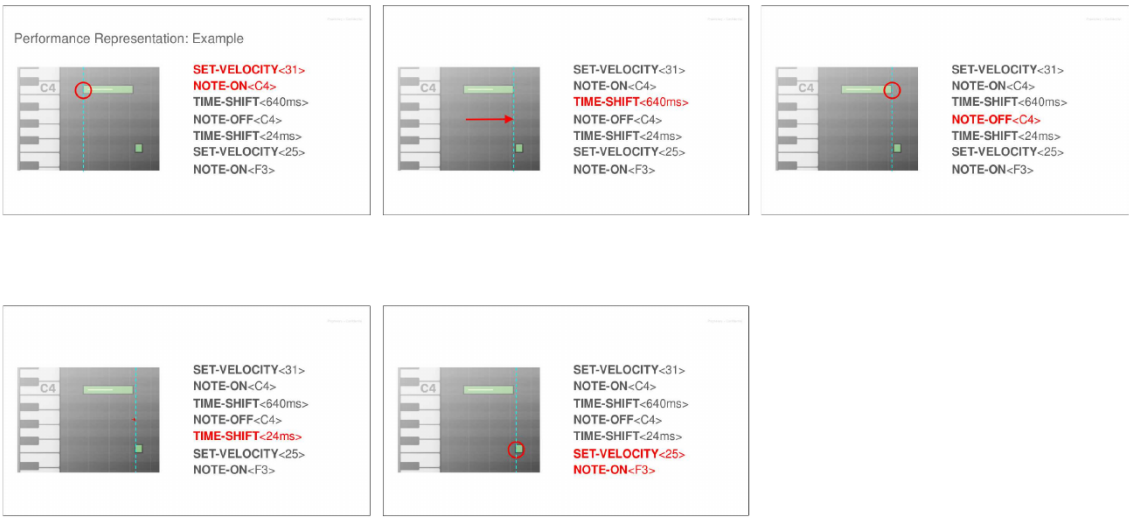


图2.2：将钢琴卷帘转换为基于事件的表示。引自 Oore 等人（2020年）。许可：CC 4.0 BY

基于事件的编码最初是为单乐器音乐开发的，尤其是钢-琴音乐（Oore 等人，2020年）。不过，只需将乐器信息注入NOTE-ON 和NOTE-OFF 词元中（Donahue 等人，2019年），即可轻松实现对乐器信息的编码。这种改动

使多乐器音乐得以表征，但会增加词汇表大小。因此，我们会得到诸如 NOTE-ON-PIANO<C4> 和NOTE-OFF-GUITAR<F3>这样的词元。

现在，我们将转向与本研究相关的下一个模态：情绪。

2.2 情绪表征

情感是支撑本论文多个子任务的核心概念。它构成了基于视频的音乐生成器的两个分支之一，充当连接视频内容与音乐输出的桥梁。在对视频进行分类时，我们将情绪用作目标和输出；在生成音乐时，则将其用作输入条件。然而，不同于视频这类始终以像素数组表示的模态，情绪可以通过多种方式建模，因此其计算建模尚未形成统一标准。

情绪表征是情感计算和心理学中的一个基本概念，是分析和建模人类情感的基础（Calvo & Mac Kim，2013年；Mat- suda 等，2013年；Morgan，2019年）。情绪表征主要有两种方法：类别模型和维度模型。

Ekman（1971年）提出了一种类别方法，使用离散类别对面部表情和生理反应进行分类，例如快乐、悲伤、愤怒、恐惧、惊讶和厌恶。维度方法则将情绪表示为连续空间中的点，通常由效价（从不愉快到愉快）和唤醒度（从平静到兴奋）定义（Russell，1980）。这两种方法各有优势与局限。类别模型更贴合人类感知和标记情绪的方式，简单明了。然而，它们依赖固定的一组离散类别，因此表征较为粗略，可能无法体现情绪体验流动且相互重叠的特性。相比之下，维度模型能够细致而灵活地表征情绪状态，通过连续的数值坐标捕捉细微变化。但这种复杂性可能使其不够直观，也更难让大众理解。

2.2.1 类别型情感表征

Ekman（1971年）对面部表情和情绪进行了开创性研究，指出六种普遍存在于不同文化中的面部情绪：快乐、悲伤、愤怒、恐惧、惊讶和厌恶。这些类别至今仍被用作现代情感相关数据集的标签（B. Xu 等人，2018年）。Izard（2013年）进一步探讨情绪在早期发展和心理功能中的作用，提出了范围更广的十种基本情绪：喜悦、兴趣、悲伤、愤怒、厌恶、轻蔑、恐惧、羞耻、内疚和惊讶。

Plutchik（1988年）提出了另一种框架，将八种基本情绪——喜悦、信任、恐惧、惊讶、悲伤、厌恶、愤怒和期待——排列成环形结构，并将它们两两配对为对立情绪，从而形成情绪轮。每种基本情绪又进一步划分为高强度和低强度的对应情绪，而次级情绪则由两种基本情绪混合产生。他对情绪轮的表征如图 2.3 所示。值得注意的是，他关于

情绪强度的概念与维度模型相符，其中唤醒度是定义情绪强度的一个关键维度。

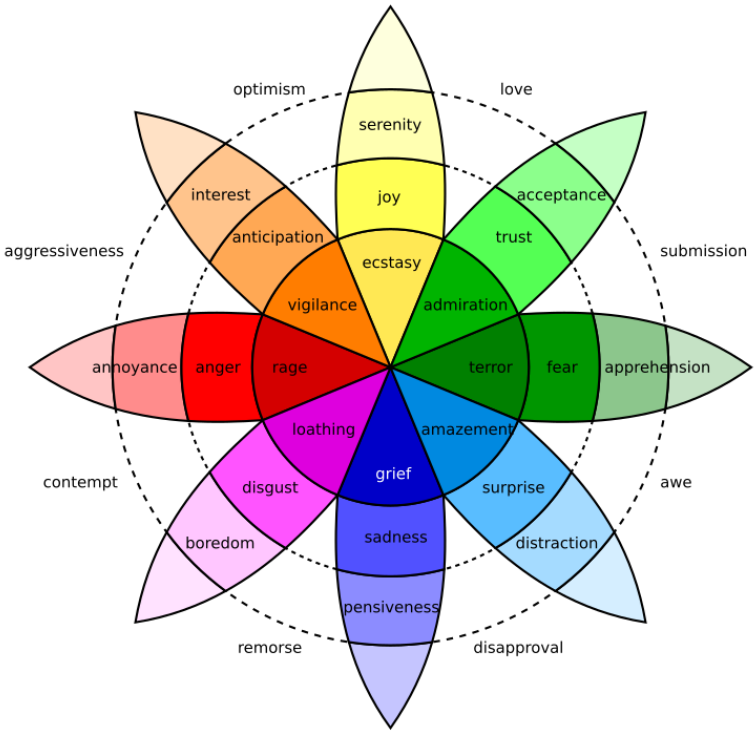


图2.3：Plutchik 的情绪轮（1988年）。作者/版权持有者：Machine Elf 1735。版权条款和许可：公有领域。

2.2.2 维度情感表征

维度模型将情感表示为连续空间中的点，而非固定的类别。这些模型利用不同轴上的数值来描述情绪，从而能够更加灵活、细致地表征情绪状态。Russell（1980）提出了著名的情感环形模型，将情绪映射到二维圆形平面上。水平轴（效价）表示情绪的愉快或不愉快程度（例如，喜悦为正面，而悲伤为负类）。垂直轴（唤醒度）表示激活程度或能量水平（例如，愤怒为正面，而平静为负类）。

由于只需两个连续值即可表示情绪，情感环形模型被广泛应用于情感分析任务，包括从图像中进行情绪预测（Savchenko，2023年）、生理信号（Wiem 与 Lachiri，2017年）以及音乐（J. Kim 等，2011年）中的情绪预测。图2.4 展示了情感环形模型，同时说明了类别情绪如何分布在其各个象限中。

介绍完关键数据模态——MIDI 和情感之后，我们现在来介绍本研究的核心架构：Transformer。

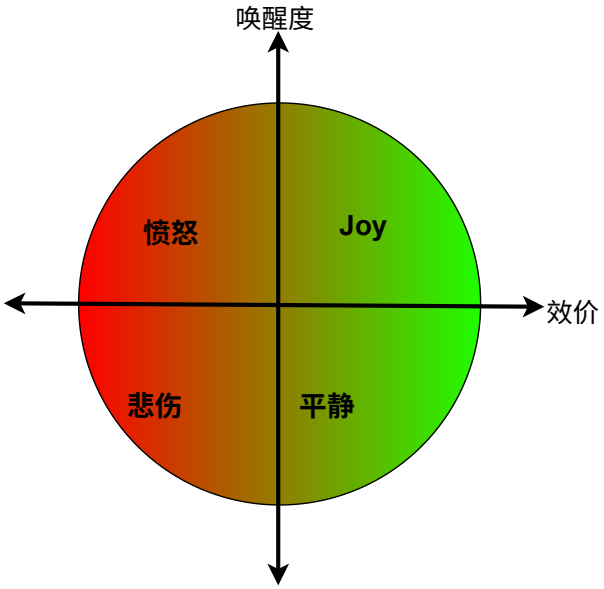


图 2.4：情绪环形模型（Russell，1980）。

2.3 Transformer

在本论文中，我们处理若干序列处理任务，并将在后续章节中详细介绍。MIDI 被视为音符序列，Transformer 则用于对这些序列进行建模和生成。在视频分类中，从各帧提取的特征构成一个序列，Transformer 可以处理该序列，以预测情绪或电影类型。我们也处理原始音频，而原始音频通常过长，无法一次性处理。为此，我们将音频拆分为数据块，从每个数据块中提取特征，然后使用 Transformer 处理生成的序列。同样，在基于文本的任务中，词或子词序列会输入 Transformer，以进行情绪分类。

Transformer 彻底革新了深度学习，尤其是自然语言处理（NLP）。它们构成了所有大语言模型的基础，包括 ChatGPT[1], Llama[2], 和 Gemini[3]。与循环模型不同，Transformer 可以并行处理整个序列，无需循环结构。借助注意力机制和并行计算，Transformer 在机器翻译（Costa-jussà 等人，2022）、文本生成（Iqbal 与 Qureshi，2022）和问答（Zaib 等人，2022）等自然语言处理任务中创下了新的性能基准。然而，它们的优势并不局限于自然语言处理。现代自动语音识别（Radford 等人，2023）、图像处理（Dosovitskiy 等人，2021）和 MIDI 生成（C. A. Huang 等人，2019年）架构也都依赖 Transformer。

现在，我们从所示框图开始，解释 Transformer 背后的理论

[1] https://chatgpt.com[2]
https://llama.com[3]
https://gemini.google.com

图 2.5。完整的架构由一个编码器（左侧）和一个解码器（右侧）组成。这种编码器—解码器架构常用于机器翻译等任务：编码器处理源序列，解码器生成目标序列。在纯生成任务中，例如文本或 MIDI 生成，只使用解码器。在这种配置中，目标序列是输入序列向右移位后的版本，这意味着对于位置 n上的输入词元，目标词元位于位置 $n+1$。简单来说，给定一个输入序列，模型会预测下一个词元。

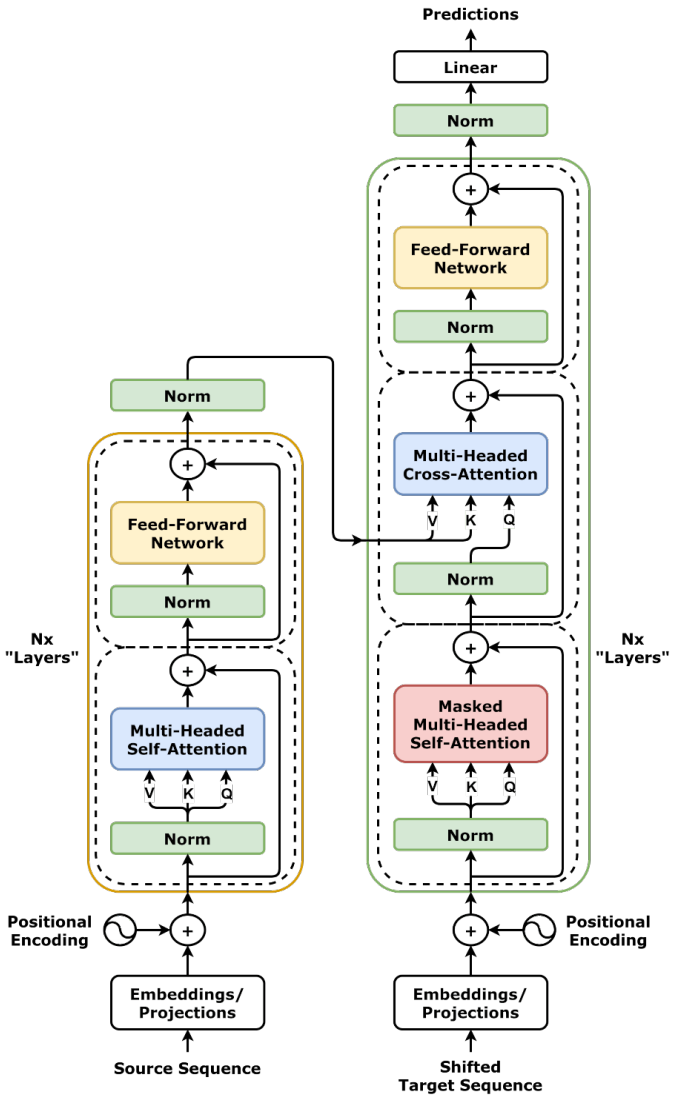


图 2.5：Transformer 模型，左侧为编码器，右侧为解码器。- https://github.com/dvgodoy/dl-visuals/?tab=readme-ov-file，CC BY 4.0

词元是序列中的离散元素。在 MIDI 处理中，词元对应单条 MIDI 消息，例如触发一个具有特定音高的音符。在文本处理中，词元通常代表词语，不过也常使用子词来捕捉词形变化之间的语义关系（Sennrich 等人，2016b）。这样可以避免对具有共同词根的词语进行冗余分词

，从而提高效率。

提取每个词元后，由所有不重复的词元构成词汇表。每个不重复的词元都可以用词汇表中的一个数值索引表示，从而将输入序列表示为一系列离散数字。嵌入是指将这些离散数字投影到具有连续值的向量中的过程。嵌入层本质上是一个查找表，定义为大小为$V \times D$的权重矩阵，其中V是词汇表长度，D是模型维度。嵌入层将每个词元的索引映射到该矩阵中对应的行，提取一个长度为D的向量。随后，沿序列维度拼接这些向量。对于长度为S的输入序列，嵌入层会生成一个大小为$S \times D$、由连续值组成的矩阵。

位置编码 与循环神经网络（RNN）不同（Rumelhart 等人，1986），Transformer 并不按顺序处理序列，而是并行处理整个输入序列，从而显著提高效率。然而，这种并行化也带来了一个关键挑战：由于自注意力会同时处理所有输入词元，Transformer 缺乏对词序的内在感知。

下面的例子展示了位置顺序的重要性。没有位置信息，以下两个句子会被理解为相同：

1- “我只是说她可以帮助他。”

2- “我说了只有她可以帮助他。”

这两个句子的区别在于单词“only”的位置，它改变了强调重点和含义。在第一个句子中，“only”修饰动词“said”。这意味着说话者的动作仅限于说她可以帮助他，并没有做其他事情。这暗示说话者除了说她可以帮忙之外，并没有表达其他意思。在第二个句子中，“only”修饰“she”。这意味着说话者强调她是唯一能帮助他的人。除了她，没人能帮忙。

为了纳入位置信息，Transformer 使用位置编码，其定义方式多种多样。原始 Transformer 使用正弦位置编码：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/D}}\right) \tag{2.1}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/D}}\right) \tag{2.2}$$

这会产生正弦波和余弦波交错排列的模式，如图所示 2.6。

通过将该位置信息简单地加到嵌入的序列中，可以将其注入模型。较新的模型使用学习式位置编码，而不是固定的正弦波（B. Wang 等，2020年）。这种学习式编码表示为一个 $S_{max} \times D$ 矩阵，其中 S_{max} 是最大输入序列长度。

C. A. Huang 等人（2019年）提出了音乐 *Transformer*，其中相对位置编码（Shaw 等人，2018年）被巧妙地融入注意力机制。与绝对

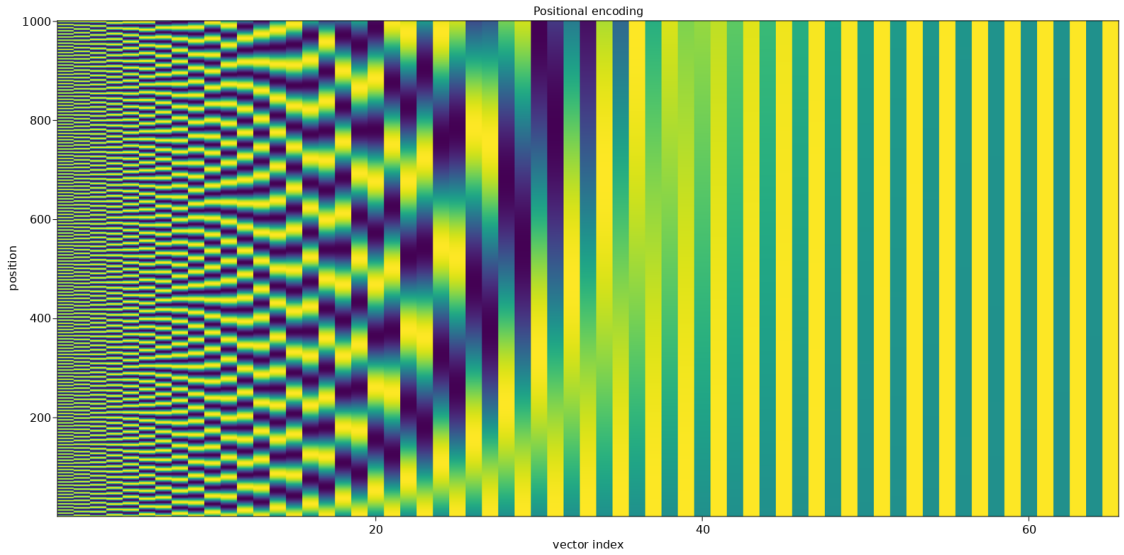


图 2.6：正弦位置编码。作者：Cosmia Nebula，个人作品，CC BY-SA 4.0，https://commons.wikimedia.org/w/index.php?curid=119948133

位置编码不同，相对位置编码表示的是词元之间的距离，而非它们的绝对位置。在音乐中，节奏与和声由音符之间的相互作用决定，而不是由它们固定的时间位置决定。因此，相对位置编码特别适用于音乐生成，因为它反映的是音乐事件之间的关系，而非它们在时间上的绝对位置。

归一化 内部激活的归一化对于任何神经网络的稳定训练都至关重要。传统神经网络通常使用批归一化，沿批次维度对激活值进行归一化（Ioffe 和 Szegedy，2015年），而 Transformer 使用层归一化，即沿特征维度对激活值进行归一化（Ba 等，2016）。

多头注意力 Transformer 的核心是注意力机制，它根据输入值的相关性计算加权和。这使 Transformer 能够同时处理句子中的所有单词，并确定每个单词相对于其他单词的重要性。

考虑将一个句子从源语言翻译成目标语言的例子。在自注意力中，模型着重理解目标句子中各单词之间的关系。在交叉注意力中，模型仍关注目标句子，同时也会考虑源句子的影响。虽然自注意力和交叉注意力背后的机制相同，但关键区别在于输入。

在多头注意力中，词嵌入首先通过单独的投影矩阵投影为查询、键和值。在自注意力中，查询、键和值均来自同一个句子。相比之下，在交叉注意力中，源序列生成键和值，而目标序列生成查询。

$$Q = X_Q W_Q, \quad K = X_K W_K, \quad V = X_V W_V \tag{2.3}$$

在上述公式中，X_Q、X_K和X_V 分别表示查询、键和值的输入词嵌入。W_Q、W_K和W_V 表示学习得到的投影矩阵。

注意力机制随后计算键与查询之间的交互，为每个值分配权重（重要性）。

$$\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_k}}\right)V \tag{2.4}$$

查询（Q）、键（K）和值（V）矩阵决定注意力输出。点积QK^T 用于计算相似度，而 $\sqrt{D_k}$（键的维度的平方根）则用于缩放这些值，以防止梯度过大。softmax 函数将分数转换为概率，并用这些概率对V中的值进行加权。

注意力通过查询（Q）、键（K）和值（V）相互比较词语的重要性。这些是输入序列经过学习得到的表征。查询（Q）表示我们要寻找的内容，例如需要结合上下文理解的词语。键（K）充当比较时的参照，例如序列中的其他词语。值（V）包含待处理的实际信息，例如我们要为其分配重要性的词语。

多头注意力不是只计算一次注意力，而是并行计算多次：

$$\text{head}^i = \text{Attention}(W_Q^i, W_K^i, W_V^i) \tag{2.5}$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W_O \tag{2.6}$$

每个注意力头（head_i）使用不同的权重对输入进行投影。所有注意力头的输出会拼接起来，并通过W_O进行线性变换。使用多个注意力头可以让模型并行捕获不同的关系。

前馈网络 图中以黄色方块表示的前馈网络2.5 会单独处理每个词，而不考虑词与词之间的交互。它通常由两个线性层组成，中间采用修正线性单元（ReLU）激活（Nair & Hinton，2010）。第一个线性层将输入投影到更高的维度，第二个线性层则将其降回模型原本的维度。这种维度扩展增强了网络的表示能力。

为稳定训练，模型中广泛使用残差连接（He 等，2016）。最后一个线性层将激活值投影回与词汇表大小相对应的维度。因此，输出向量表示词汇表中每个词的概率。

介绍了关键概念后，我们现在转而回顾与本文所探讨任务相关的文献。

2.4 序列处理

符号音乐可以表示为序列数据，与文本类似。因此，原则上，同一模型既可用于自然语言处理（NLP），也可用于符号音乐处理。用于序列建模的最早一类神经网络架构是循环神经网络（RNN），每个时间步都会处理一个输入样本（词元）（Rumelhart 等人，1986）。网络通过一种名为时间反向传播的算法，计算各时间步的误差梯度来进行训练。

然而，RNN 在建模长序列方面并不十分成功，因为随着序列变长，反向传播的梯度可能趋近于零。这一问题称为梯度消失问题。长短期记忆网络（LSTM）通过使用专门的门控缓解了这一问题（Hochreiter 和 Schmidhuber，1997年）。一种类似的变体——门控循环单元（GRU）——能够取得与 LSTM 相近的演奏效果，同时架构更简单、参数更少（Cho、van Merrienboer 等人，2014b）。

尽管有这些改进，LSTM 和 GRU 在高效处理长序列方面仍然面临困难，并且随着序列变长，往往会“忘记”较早的输入样本。注意力机制通过显式建模所有输入样本对之间的依赖关系来解决这一问题（Bahdanau 等人，2015年）。最后，Transformer 模型通过在多头、多层架构中融入注意力机制，在序列处理方面达到了当前最先进水平（Vaswani 等人，2017年）。

最初的 Transformer 实现由一个编码器网络和一个解码器网络组成，并在机器翻译任务上进行了测试。编码器—解码器架构常用于机器翻译，其中编码器处理源文本，解码器生成输出（Cho、van Merrienboer 等人，2014a）。由于语言建模涉及从头开始生成文本，因此可以将其视为与音乐生成类似，这意味着这两项任务都可归类为序列生成任务。

在序列生成中，不存在彼此独立的源序列和目标序列。因此，最先进的语言模型仅由解码器构成（Brown 等人，2020年）。序列生成神经网络使用来自同一领域的输入序列和目标序列进行训练。具体来说，目标序列是输入序列的移位版本，因此，对于每个输入词元，神经网络都会预测下一个词元。

在序列生成模型的训练过程中，训练损失根据所有词元的预测性能计算，即使模型只考虑前面的词元。为确保模型不会关注未来词元，需要在注意力模块的 Softmax 输出中加入三角掩码。下面展示了序列长度为 5 时的三角掩码示例。

$$mask = \begin{bmatrix} 0 & -\infty & -\infty & -\infty & -\infty \\ 0 & 0 & -\infty & -\infty & -\infty \\ 0 & 0 & 0 & -\infty & -\infty \\ 0 & 0 & 0 & 0 & -\infty \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

这样可以确保与未来词元相关的注意力输出被掩码。例如，当模型预测第 3 个词元时，它只会使用第 1 个和第 2 个词元。此生成任务可以视为针对所有词元的分类任务。此任务最常用的损失函数是交叉熵损失：

$$\mathscr{L} = -\sum_{s=1}^{S}\sum_{v=1}^{V} y_{s,i}\log \hat{y}_{s,i} \tag{2.7}$$

该损失函数用于衡量预测词元概率与真实标签之间的差异。索引s 表示序列中的位置，范围从 1 到 S，其中 S 是词元总数。索引 i 对应词汇表中的词语索引，范围从 1 到 V，其中 V 是词汇表大小。术语 $y_{s,v}$ 是独热编码标签，这意味着如果位置s 上的正确词语是词语 v，则其值为 1，否则为 0。术语 $\hat{}y_{s,v}$ 是位置s上词语i 的预测概率，该概率由 Transformer 模型的 Softmax 输出得出。对数 $\log \hat{}y_{s,v}$ 确保错误预测会受到严厉惩罚，因为预测概率越低，负值就越大。由于每个位置只有一个正确词语，因此在每个时间步的求和中，只有一项会对损失产生贡献。模型通过最小化 $\mathscr{L}$ 来优化此函数，从而提高正确词语的概率，同时降低错误预测的可能性。

序列生成模型在推理过程中以自回归方式运行，这意味着每次只生成一个新词元。生成每个词元时，都会执行一次前向传播。生成词元后，会将其添加到序列中，再将更新后的序列输入模型，以预测下一个词元。例如，如果我们已经生成了由词元 1 和 2 组成的序列，就会将该序列作为输入来预测词元 3。词元 3 生成后，会将其追加到序列中，形成序列 [1, 2, 3], ，再用它来预测词元 4，依此类推。通常会使用一个特殊的 开始词元作为第一个词元，以从头开始进行推理。或者，也可以使用预先定义的序列进行推理，这种序列称为提示序列或提示序列。在这种情况下，模型会从用户提供的序列继续生成，生成的词元将扩展该序列。

与生成任务不同，分类任务会针对整个序列预测一组概率，而不是为每个词元分别预测。为此，输入序列前会添加一个名为 `CLS`（分类）的特殊词元，并且只考虑与该词元对应的输出。在分类任务中，不使用三角掩码，因此每个词元都可以关注其他所有词元。因此，与 `CLS` 词元对应的输出会利用整个序列的信息。

在接下来的章节中，我们将讨论与本研究相关的序列分类和序列生成任务。首先回顾文本情感分类，随后介绍条件序列生成，特别是条件音乐生成。

2.4.1 文本情感分类

文本情感分类——在机器学习文献中也常与情感分析互换使用——能够自动识别和/或量化文本片段中表达的情绪，例如评论、社交媒体帖子或客户反馈（B. Liu & Zhang, 2012年）。识别文本中潜在的情绪在多个领域都很有用，例如客户服务（M. Hu & Liu, 2004年）、金融（Nguyen 等人，2015年）、政治（Iyyer 等人，2014年）和娱乐（J. Almeida 等人，2021）。

现在，我们将介绍最新且最全面的文本情感分类数据集及相关方法。

2.4.1.1 数据集

SemEval（语义评测）研讨会定期举办各种聚焦于语义文本分析的挑战赛（Ojha 等人，2024）。其中一些挑战赛旨在开展基于文本的情绪分类。2025年届的任务11名为“弥合基于文本的情感检测鸿沟”，汇集了来自社交媒体、演讲、文学和新闻等不同来源的样本（Muhammad 等人，2025年）。该数据集包含来自30多种语言的超过100,000个样本，所有样本均由人工标注为 Ekman 的六种基本情绪——喜悦、悲伤、愤怒、恐惧、惊讶和厌恶——以及中性类别（Ekman，1971年）。其中，英语样本有5,651个。2019年届的任务3——“EmoContext：文本中的上下文情感检测”——要求参与者判定一段简短对话中最后一句话所表达的情绪（Chatterjee 等人，2019年）。该数据集包含38,424个样本，标注为四类：开心、悲伤、愤怒和其他。2018年届的任务1名为“推文中的情感”，聚焦于 Twitter 数据中的情感内容[4]（Mohammad 等人，2018年）。其中包含多个子任务：情绪强度回归、强度序数分类、效价回归、效价序数分类以及情绪分类。该挑战赛支持英语、西班牙语和阿拉伯语三种语言，其中专门用于情绪分类任务的英语样本有10,983个。

EmoBank 数据集包含 10,062 个英语句子，收集自新闻文章、博客、小说和旅行指南等来源。这些样本通过众包方式，采用维度情绪表示方案进行标注（Buechel 与 Hahn，2017年）。多模态 EmotionLines 数据集（MELD）旨在识别多方对话中的情绪（Poria 等人，2019年）。该数据集包含文本、视频和音频数据。其文本部分包含 13,000 条英语话语，并依据 Ekman 的六种基本情绪——喜悦、悲伤、愤怒、恐惧、惊讶和厌恶——以及中性类别进行标注（Ekman，1971年）。清洗平衡情感推文（CBET）数据集提供了 81,163 条推文，并标注了九种情绪：喜悦、愤怒、

[4] https://twitter.com

悲伤、爱、轻蔑、惊讶、恐惧、内疚和感激（Shahraki，2015年）。GoEmotions 数据集包含 58,000 条来自 Reddit[5] 评论的英语文本样本（Demszky 等人，2020年）。该数据集采用涵盖 27 种类别情感的丰富标签体系，包括钦佩、愉悦、认可、恼怒、愤怒、好奇、困惑、关怀、渴望、厌恶、不赞同、失望、尴尬、兴奋、恐惧、喜悦、悲痛、感激、爱、悲伤、紧张、乐观、自豪、懊悔、宽慰、恍然大悟和惊讶。

表 2.1 总结了上述提到的文本情绪数据集。该表概述了每个数据集的来源或上下文、英语样本数量，以及所采用的情感表征类型。

表 2.1：文本情感分类数据集

数据集	上下文	样本量	情感表征
CBET (Shahraki，2015年)	推文	81,163	类别型
EmoBank (Buechel 与 Hahn，2017年)	新闻、博客、小说、旅行指南	10,062	维度
SemEval-2018 (Mohammad 等人，2018年)	推文	10,983	类别型
SemEval-2019 (Chatterjee 等人，2019年)	对话	38,424	类别型
MELD (Poria 等人，2019年)	对话	13,000	类别型
GoEmotions (Demszky 等人，2020年)	Reddit 评论	58,000	类别型
SemEval-2025 (Muhammad 等人，2025年)	社交媒体、演讲、文学、新闻	5651	类别型

2.4.1.2 方法

在过去二十年里，机器学习方法显著推动了文本情感分2002年类领域当前最先进水平的发展。早期研究依赖手工设计特征，例如常用的 n 元语法（Pang 等人，2005年），或与特定情绪相关的形容词和副词（Whitelaw 等人，2012年）。不过，深度学习的出现使得无需手动提取特征、直接处理原始输入在计算上变得可行，从而提升了演奏表现（Krizhevsky 等人，1986）。最初使用的是循环神经网络及其改进变体，如长短期记忆网络（Rumelhart 等人，2017年），但后来被 Transformer 模型取代（Vaswani 等人，），该模型是当前文本分类任务中最先进的技术。2025 年 SemEval 文本情感分类任务以“弥合基于文本的情绪检测鸿沟”为题，指出当前最先进的方法依赖于经过微调的预训练 Transformer，或经由指令微调、适配器模块或数据增强技术增强的大语言模型（LLM）

[5] https://www.reddit.com

（Muhammad 等人，2025年）。

针对特定任务对预训练 Transformer 进行微调，可以利用模型在初始训练阶段获得的知识。例如，GPT（生成式预训练 Transformer）模型是一种大型 Transformer，经过下一词元预测预训练，随后针对各种自然语言处理任务进行微调，从而达到最先进的性能（Radford 等人，2018年）。BERT（来自 Transformer 的双向编码器表征）通过在预训练期间引入掩码词元预测，改进了这一方法，使模型能够学习双向上下文（Devlin 等人，2019年）。在 BERT 的基础上，RoBERTa（稳健优化的 BERT 方法）模型通过广泛的超参数调优、使用更多数据训练、采用更大的批次大小以及延长预训练时值，进一步提升了性能（Y. Liu 等人，2019年）。

指令微调是指使用指令对大语言模型（LLM）进行监督式微调，指令通常采用提示-输出对的形式（S. Zhang 等人，2023年）。由于对整个大语言模型进行微调可能需要大量资源，许多方法会使用适配器来减少可训练参数的数量（Houlsby 等人，2019年）。适配器是插入神经网络内部的小型模块，可以在冻结大型网络权重的同时对模型进行微调。这样既能训练适配器模块，又能保持较高的计算效率。低秩适配器（LoRA）等先进适配器使用低秩矩阵分解，以比原始权重矩阵更少的参数表示权重更新（E. J. Hu 等人，2022年）。Leon 等人（2025年）利用多个语言的数据集进行跨语言文本情感分类。Ranjbar 和 Baghbani（2025年）在微调预训练 Transformer 时使用大语言模型进行数据增强。他们使用 LLaMa-3 模型[6] 为训练数据集中的文本生成解释，从而提升了预训练 RoBERTa 模型在文本情感分类任务上的性能。

接下来，我们将介绍序列处理的背景知识，重点关注条件序列生成。

2.4.2 条件通用序列生成

尽管这一术语的用法较为宽泛，条件化 指的是通过提供辅助输入（即条件）来控制模型的输出。这些条件可以与输入和目标属于同一领域，从而能够利用无标签数据进行训练。或者，条件也可以属于不同领域，例如带标签数据集的标签。

早期的自然语言处理神经网络就已采用条件化。Mikolov 和 Zweig（2012年）开发了一种根据主题和类型等因素进行条件化的语言模型，其中，条件向量由线性层生成，并与循环神经网络（RNN）的隐藏状态拼接。Sennrich 等人（2016a）开发了一种用于英语到德语翻译的编码器—解码器 RNN 模型，并根据礼貌程度进行条件化。他们使用控制标记来指定用户偏好正式翻译还是非正式翻译。

[6] https://llama.com

条件变换器语言（CTRL）模型将用于标示领域、风格、主题等内容的控制标记输入大型 Transformer，在条件语言建模方面取得了最先进的结果（Keskar 等人，2019年）。Krishna 等人（2020年）通过生成释义文本进行风格迁移，并证明针对每种风格分别训练模型，优于训练一个使用特定风格控制标记的单一模型。

Sheng 等人（2020年）识别出触发序列——作为输入时会生成有偏文本的子序列——并将其用作提示序列，以在语言建模中诱发或平衡偏见。Smith 等人（2020年）通过比较三种方法，研究了如何控制对话生成的风格：检索与精炼（Weston 等人，2018年）、推理时迭代精炼（Dathathri 等人，2020年）以及使用控制标记进行条件生成（Keskar 等人，2019年）。他们表明，使用控制标记进行条件生成的效果优于其他方法。

文献中的大多数研究使用类别变量（例如控制标记）来引导语言建模。相比之下，图像描述可以被视为一种基于图像的文本生成任务，而图像属于非类别型变量。在此情境下，输入图像通常由卷积神经网络（CNN）处理，得到的特征则用于对独立的语言模型进行条件化。早期研究在这一任务中使用循环神经网络（RNN）作为语言模型（Vinyals 等人，2017年），但最先进水平的模型已用 Transformer 取代 RNN（X. Zhu 等人，2018年）。X. Zhu 等人（2018年）比较了不同的条件化方法，发现它们的演奏效果相近。这些方法包括将图像空间特征输入解码器的交叉注意力层（K. Xu 等人，2015年）、将图像特征与每个词嵌入结合，以及在词嵌入之前输入图像特征（Vinyals 等人，2017年）。

接下来，我们将讨论与整体工作最密切相关的任务：条件音乐生成。

2.4.3 条件符号音乐生成

虽然我们将重点放在符号音乐的条件生成上，但同样有必要了解符号音乐的非条件生成。接下来，我们将介绍相关的数据集和方法。

2.4.3.1 数据集

规模最大的符号音乐数据集是 Lakh MIDI 数据集（LMD）（Raffel，2016），其中包含 176,581 个未标注的多乐器 MIDI 文件。在这些文件中，有 45,129 个与百万歌曲数据集（MSD）中的 31,034 首曲目相匹配（Bertin-Mahieux 等人，2011年）。数量不一致的原因是，多个 MIDI 文件可能对应同一首 MSD 曲目，而单个 MIDI 文件也可能对应多首曲目。

MAESTRO 数据集（MIDI 与同步音轨和组织音频编辑）是一个大型、高质量的数据集，旨在支持音乐研究，尤其是转录、生成和合成等领域（Hawthorne 等人，2019年）。其中包含 200 多小时的专业

钢琴演奏音频，并配有精确对齐的音频和 MIDI 录音。MIDI 数据不仅记录音符和时序，还记录音符力度，用以表示每次按键的强度或力度变化；这对于建模富有表现力的钢琴演奏至关重要。

Zhuo 等人（2023年）介绍了符号音乐视频（SymMV）数据集，其中包含视频-MIDI 配对。他们从钢琴教学视频中获取 MIDI 数据，并使用 Onsets and Frames 模型自动转录音频（Hawthorne 等人，2018年），最终得到 1,140 个样本。他们还提供了低层特征，例如颜色直方图和用于表示运动的 RGB 帧差，以及从视频中提取的高级 CLIP 特征。

据我们所知，目前只有四个公开可用的带有情绪标签的符号音乐数据集，但它们的样本量都很小。VGMIDI（电子游戏 MIDI）数据集包含 204 首用钢琴演奏的电子游戏配乐，其效价和唤醒度采用连续值标签，由 30 名人类受试者标注（Ferreira 和 Whitehead，2019年）。Panda 等人（2013年）创建了类 MIREX 数据集，其离散情绪标签取自 AllMusic 数据库[7]。该数据集主要由音频录音组成，但其中 193 个样本也提供 MIDI 文件，且乐器数量各不相同。EMOPIA（情绪钢琴）数据集包含钢琴独奏的配对音频和 MIDI 片段，取自 387 首歌曲（Hung 等人，2021）。作者根据情感二维环形模型的四个象限，采用离散类别对情感标签进行了人工标注（Russell，1980）。表 2.2列出了 MIDI 数据集、总时值（小时）、乐器编制以及情绪标签的类型。

遗憾的是，由于规模较小，这些数据集不足以训练具有数百万参数的深度神经网络，因为它们往往会导致过拟合和较差的泛化能力。总体而言，目前还没有标注情绪标签的大规模多乐器 MIDI 数据集。

表 2.2：MIDI 数据集

数据集	时值（小时）	仪器设备	情绪标签
类似 MIREX (Panda等人，2013年)	11.6	多乐器	类别型
LMD (Raffel, 2016)	9382.1	多乐器	None
MAESTRO (Hawthorne 等人，2019年)	200.0	仅钢琴	None
VGMIDI (Ferreira 和 Whitehead，2019年)	6.4	仅钢琴	维度
EMOPIA (Hung 等人，2021)	11.0	仅钢琴	类别型
SymMV (Zhuo 等人，2023年)	76.5	仅钢琴	None

[7]https://allmusic.com

2.4.3.2 方法

早期使用神经网络进行音乐生成的研究采用了循环神经网络（RNN）（Eck & Schmidhuber，2002年）。然而，近年来 Transformer 模型的出现使对更长依赖关系的建模成为可能。音乐 Transformer 通过融入相对位置信息，在 Transformer 模型的基础上进行了构建，从而更好地表征音符之间的关系（C. A. Huang 等人，2019年）。

现有关于符号音乐生成的文学大多依赖非条件方法。这些方法在没有显式标签的原始 MIDI 数据上进行训练，从而能够生成与训练数据集中示例相似的新音乐（C. A. Huang 等人，2019年）。然而，一些方法利用数据中的低层特征，以条件化方式生成音乐（C. A. Huang 等人，2017年）。例如，它们可以使用短旋律、和弦或单乐器音轨作为输入，以生成相应的旋律。使用短旋律作为输入特别实用，因为它们被直接输入模型，无需改变架构。在序列处理中，这些短输入序列被称为引导序列。随后，模型预测紧随提示序列旋律之后的旋律。在这种情况下，条件和目标都属于同一领域，因此可以使用无标签数据训练模型。

其他利用同领域条件化的符号音乐生成任务包括伴奏生成（Dong 等人，2018年；C. A. Huang 等人，2017年）、插值（Roberts 等人，2018年）、修复（K. Chen 等人，2020年；Ens & Pasquier，2020年）以及风格迁移（Choi 等人，2020年；Z. Wang 等人，2020年）。MidiNet 可通过在包含和弦信息的私有数据集上训练，生成以和弦为条件的旋律（L. Yang 等人，2017年）。OpenAI 的 MuseNet 在多个数据集的组合上训练，并且可以使用引导序列控制标记，使生成以特定艺术家姓名、类型或风格为条件（Payne，2019年）。

还可以使用低层符号音乐特征进行条件化（Guo 等人，2020年；Park & Kim，2000a；R. Yang 等人，2019年）。速度、音符密度、音高范围和调性张力等特征可以自动计算，从而无需使用带标签数据集。H. H. Tan 和 Herremans（2020年）旨在通过使用无标签的 MAESTRO（用于同步音轨与组织的MIDI和音频）数据集（Hawthorne 等人，2019年）来扩充规模较小的带标签 VGMIDI 数据集——该数据集将在下一节详细介绍。他们使用低层节奏和音符密度特征来推断高层唤醒度特征。类似地，J. Zhu 等人（2024）使用附加标记纳入和弦信息，并通过生成对抗网络（Goodfellow 等人，2020年）进一步提升输出质量。

基于情绪的音乐生成，或称情感算法作曲（AAC）方法，主要关注基于特定情绪自动创作音乐（Williams、Kirke、Miranda 等，2015年）。AAC 的应用场景包括为视频和电子游戏创作配乐（Williams、Kirke、Eaton 等，2015年）、用于医疗应用的神经反馈训练（Pratt 等，1995年），以及开发脑机音乐接口（Miranda 等，2011年）。

尽管音乐与情感之间的关系已得到充分研究（Juslin & Slobod a,

2001），但选择哪种“最优”情绪模型来探讨这种关系，仍是一个存在争议的主题（Eerola & Vuoskoski，2011年）。现有的 AAC 研究主要采用情绪模型的两大类别：类别型和维度型（Williams、Kirke、Miranda 等，2015年）。

正如第2.2.1节所讨论的，分类情绪模型使用离散标签，如快乐、悲伤、愤怒和惊讶（Ekman，1975）。然而，关于维度方法的研究认为，类别模型无法捕捉人类情感的复杂性和微妙之处。相反，他们建议使用连续值坐标在低维空间中表示情绪（Gunesetal.，2011年）。

AAC的早期研究使用各种旋律、和声和节奏特征来针对特定情绪（参见Williams、Kirke、Miranda等人（2015年）的综述）。然而，这是一种间接方法，情绪与音乐特征之间的对应关系仅是近似的（Berg与Wingstedt，2005年；Wingstedt等人，2005年）。深度学习的出现使得可以使用在大型带标签数据集上训练的复杂模型，直接利用情绪标签，从而无需中间特征（Krizhevsky等人，2012年）。

较新的 AAC 模型采用端到端方式，在包含符号音乐和情绪标签的数据集上进行训练。VGMIDI 数据集的创建者提出了一种以情感为条件的符号音乐生成方法（Ferreira & Whitehead，2019年）。他们使用遗传算法，分别针对正面和负面效价条件，对预训练长短期记忆网络的权重进行微调，最终得到两个模型。

K. Zhao 等人（2019年）和 Hung 等人（2021）都生成了以四种类别情绪为条件的符号音乐，每种情绪分别对应效价-唤醒平面的四个象限之一。K. Zhao 等人（2019年）使用类别情绪标签对从 piano-midi.de 获取的古典钢琴音乐进行标注，并在这个带标签数据集上训练了双轴长短期记忆网络（Johnson，2017年）。EMOPIA 数据集的创建者 Hung 等人（2021）训练了一个以控制标记为条件的 Transformer 模型（Keskar 等人，2019年）。然而，这些研究受到可使用的类别情绪标签数量较少的限制，这很可能是由于其训练数据集规模较小。

我们将在第2.6 节讨论基于视频的符号音乐生成，这项任务需要更复杂的模型和视频分析模块。接下来，我们将讨论视频分析任务。

2.5 视频分析

基于视频的音乐生成模型通常由视频分析模块和音乐生成器组成（Di等人，2021；Kang 等人，2024）。如第1章所述，我们的工作涉及低层次视频分析，具体而言是场景切换检测，以及高层次视频分析，即语义视频分类。

标准视频处理库（例如 FFMpeg[8]）内置了场景切换检测功能。FFm- peg 的算法会计算连续帧之间的平均像素差异，并在该差异超过指定阈值时标记场景切换。由于场景切换检测已经是一个解决了的问题

[8]https://ffmpeg.com

因此，我们将讨论重点放在语义视频分类上，尤其关注情感和流派分类。

不同于其他信号的分类，视频分类是一项多模态任务。视频分类模型同时利用视频帧（捕捉场景、物体、动作、人脸和文本）以及音频，音频包括语音、音乐、音效和音频事件。此外，由于视频是序列，还必须考虑短期与长期依赖关系。视频内容的性质也会显著影响最相关的模态。例如，在基于面部表情对情感进行分类时，来自视频帧的视觉信息至关重要。相比之下，在检测社交互动中的情感时，分析语音同样重要，甚至更为重要。

2.5.1 视频情感分类

视频情感分类是指使用计算方法，自动检测并分类视频中表达的情绪。视频分类模型以视频帧作为输入，并且通常也包含音频，输出形式为预定义情感类别的概率，或效价和唤醒度等连续值。虽然我们的工作侧重于任意视频的分类，但我们还列出了以下视频情感数据集，其中既包括任意类型的视频，也包括特定类型的视频。

2.5.1.1 数据集

Gao 等人（2021）提出了成对情感关系识别（PERR）数据集，其中包含来自电视剧的 31,182 个视频，重点关注人类之间的社交互动。上下文感知情感识别（CAER）数据集同样收集了来自 79 部电视剧的 13,201 个片段，但专门选取了包含人脸的片段（Lee 等人，2019年）。Pandeya 和 Lee（2021）为 3,438 部音乐视频标注情绪，创建了音乐视频情感数据集（MVED）。

用于任意视频分类的数据集有限，即这些数据集包含多样化的视频。VideoEmotion-8 数据集由从 YouTube 和 Flickr 收集的 1,001 个用户生成视频组成，并标注了 8 种类别情绪：愤怒、期待、厌恶、恐惧、喜悦、悲伤、惊讶和信任（Jiang 等人，2014年）。Ekman-6 数据集（B. Xu 等人，2018年）是 VideoEmotion-8 的扩展版本，包含 1,637 个视频，并按照 Ekman（1971年）定义的六种基本情绪进行标注：愤怒、厌恶、恐惧、喜悦、悲伤和惊讶。表2.3 总结了视频情感数据集，这些数据集均采用类别式情绪表征。

现在我们继续讨论视频情感分类的方法。

2.5.1.2 方法

所有现代视频分类模型都依赖于深度神经网络。B. Xu 等人（2018年）采用了跨领域知识迁移，特别是从图像情感分类到视频情感分类。C. Chen 等人（2016）首先对视频中的场景、物体和事件进行分类，并

表 2.3：视频情感数据集。

数据集	上下文	样本量
VideoEmotion-8（Jiang 等人，2014年）	用户生成	1,001
Ekman-6（B. Xu 等人，2018年）	用户生成	1,637
CAER（Lee 等人，2019年）	人脸	13,201
PERR（Gao 等人，2021）	社交互动	31,182
MVED（Pandeya 和 Lee，2021）	音乐视频	3,438

然后融合这些信息以对情感进行分类。S. Zhao 等人（2020年）使用 3D 卷积神经网络处理视频帧，并使用 2D 卷积神经网络处理音频频谱图。他们还使用了经过修改的损失函数，当预测与目标具有相似的情感极性时，对错误分类施加更高的惩罚。

Z. Zhang 等人（2023年）引入了交叉与内部注意力模块，用于分析视频帧之间以及视频帧与音频之间的依赖关系。根据注意力模块的 softmax 输出，他们识别出主要特征并将其删除，促使模型关注次要信息。

由于视频帧序列较长，许多方法会提取并处理关键帧。提取关键帧的一种直接方法是识别场景边界，而 Wei 等人（2021）则进一步根据这些帧的情感显著性进行筛选。他们将帧输入图像情感分类器，并使用概率最高的类别来计算情感显著性。本质上，他们选择能产生更显著情绪分类结果的视频帧，并在最终分析中使用这些帧。Yi 等人（2024）通过线性插值混合帧，或剪切并粘贴不同片段，采用数据增强技术进行视频情感分类。上述方法大多依赖于以数据为导向的方法，例如数据选择（如识别关键帧）或数据增强（如生成合成帧）。

现在，我们将讨论转向另一项任务，继续探讨语义视频分类：电影预告片类型分类。得益于 IMDb 等在线电影数据库以及 YouTube 等视频平台，电影预告片数据集通常样本量更大、视频时长更长。丰富的数据使研究人员能够将重点从以数据为导向的策略转向以模型为导向的方法。

2.5.2 预告片类型分类

在预告片类型分类中，模型会处理电影预告片，以预测电影的类型。这通常是一项多标签分类任务，因为电影往往同时属于多种类型。接下来，我们将继续讨论电影数据集和预告片类型分类方法。

2.5.2.1 数据集

MovieLens 是最早的电影数据集之一，其中包含电影评分以及类型、片名等元数据（Harper & Konstan，2016）。虽然最新版本收录了 8.6 万部电影，但

并未提供电影预告片的直接链接。只能通过爬取互联网电影资料库（IMDb）网站获取这些预告片[9]，这与其使用条款相冲突。

大型电影预告片数据集（LMTD）收录了 3,500 部电影预告片的特征和元数据，但该项目目前已停止，数据也不再提供（Simoes 等，2016）。多模态 IMDb 数据集（MM-IMDb）将电影海报与来自 MovieLens 的元数据相结合，包含 2.6 万部电影及其海报、剧情简介、类型和其他元数据，但不包含预告片或视频（Ovalle 等，2017年）。

多模态电影预告片特征数据集（MMTF-14K）提供从 1.4 万部电影预告片中提取的多模态特征，以及用户评论和类型等元数据（Deldjoo 等，2018年）。MovieScope 是一个用于多模态电影分析的综合数据集，包含电影预告片、海报、剧情梗概、用户评论和视听特征等数据，涵盖 5,000 部不同的电影（Cascante-Bonilla 等，2019年）。

浓缩电影数据集包含来自4000部电影的完整独立场景（而非预告片），以及剧情细节和角色信息（Bain 等人，2020年）。MovieNet 是一个大规模的整体数据集，提供电影、预告片、海报、字幕、剧情、标签和类型元数据（Q. Huang 等人，2020年）。尽管它包含37.5万部电影的元数据，但可通过 YouTube 链接获取3.3万部电影的预告片，这使得 MovieNet 成为就已标注预告片而言规模最大的电影数据集。

在预告片类型分类的上下文中，早期数据集要么不包含预告片，要么需要网页抓取，而这可能与在线平台的使用条款相冲突。较新的工作为其数据集中的电影提供了预告片的 YouTube 链接，但这些数据集的规模仍然有限。在现有选择中，MovieNet 作为提供电影预告片直接访问的最大数据集脱颖而出，使其成为训练用于预告片类型分类的深度神经网络的最合适候选。

2.5.2.2 方法

早期电影类型分类研究之一使用场景检测器获取关键帧，并从私下收集的数据集中提取手工设计的视觉特征，例如粗糙度、崎岖度和开阔度（H. Zhou 等，2010）。通过比较训练集与测试集中的特征向量之间的距离，实现了流派分类。

最早将神经网络用于电影类型分类的研究之一，也采用了视觉预训练特征（Wehrmann & Barros，2017年），并使用 LMTD 数据集（Simoes 等，2016）。研究人员采用了在 ImageNet（Deng 等，2009）和 Places365（B. Zhou、Lapedriza 等，2018年）数据集上训练的分类模型所提取的预训练特征，以及音频频谱图。研究人员使用时间卷积模块融合各帧的特征。该模块的运作方式与标准卷积神经网络（CNN）相同，其卷积核长度与每个关键帧的输入特征长度相匹配。卷积核沿时间维度遍历后续帧的特征（Wehrmann & Barros，2017年）。沿时间维度的卷积核长度

[9] https://www.imdb.com

为 3，这意味着模型仅利用 3 帧之间的对应关系。此外，由于 CNN 的输入帧数与输出帧数不一致，模型会沿时间维度取最大值，以生成供最终分类层使用的固定长度向量。这种方法不可避免地会造成信息损失。另一项研究则探索了面部情绪与电影类型之间的对应关系：首先从预告片视频中提取人脸并对其情绪进行分类，然后将这些情绪映射到电影类型（Yadav & Vishwakarma，2020年）。

介绍 MovieNet 数据集的研究还提出了一种电影类型分类模型，并报告了其他最先进视频分类模型的结果（Q. Huang 等人，2020年）。尽管他们提出的模型性能优于现有模型，但它只使用了 8 个片段，每个片段从整个预告片中选取 3 帧。推理时，模型会对每个片段分别进行预测，再取平均值以生成最终预测。MovieCLIP 模型首先训练一个场景分类模型，然后将其最终激活值作为类型分类模型的输入，具体使用的是 Moviescope 数据集（Bose 等人，2023年）。

近期一些研究使用 Transformer 进行电影类型分类。Rodriguez Bribiesca 等人（2021）使用 Transformer 分别处理预训练帧特征，然后将生成的激活值与元数据和电影海报融合。Miyazawa 等人（2022年）和 Ak 等人（2023年）使用预训练 Transformer 处理电影海报和剧情简介，不使用任何视频数据，在 MM-IMDb 数据集上进行类型分类（Ovalle 等人，2017年）。

上述所有方法都对输入数据施加了限制——要么将输入帧的数量限制在一个固定且较小的集合内，要么依赖于从标准图像分类模型（如 VGG（视觉几何组））提取的预训练帧特征（Simonyan & Zisserman，2015年）。前一种方法无法捕捉预告片中存在的长期依赖关系，而后一种方法则将模型限制在从单一上下文中提取的特征，忽略了人脸或屏幕文字等其他信息线索。这些限制凸显了对改进方法的需求：此类方法能够利用数量大且可变的帧，并融合来自多种任务的预训练特征。

接下来，我们讨论将视频分析与符号音乐生成相结合的研究，由此产生基于视频的符号音乐生成器。

2.6 基于视频的符号音乐生成

基于视频的符号音乐生成器会将输入视频作为输入，并以符号格式创作合适的配乐。多项研究聚焦于为特定类型的视频生成符号音乐，例如展现舞蹈等人体动作或器乐演奏的视频。Foley Music 模型（Gan 等人，2020年）通过图卷积网络（Kipf & Welling，2017年）和 Transformer（Vaswani 等人，2017年）处理身体关键点运动，从音乐家的视频中生成 MIDI。2020年Koepke 等人和 Su 等人（2020年）也采用带有残差连接的深度神经网络，根据钢琴键盘上的手指动作视频生成符号音乐。

由于这些方法具有专门性，因此依赖包含视频-MIDI 配对的数据集，不过这类数据集通常少于 1,000 个样本（Koepke 等人，2020年；B. Li 等人，2019年；Su 等人，2020年；H. Zhao 等人，2018年）。RhythmicNet 模型采用多阶段流程，根据舞蹈视频生成音乐：预测节拍和风格、生成鼓轨，随后创作多轨音乐（Su 等人，2021）。接下来，我们介绍解决更一般任务的研究，即为任意视频生成符号音乐。

可控音乐 Transformer（CMT）根据视频特征生成音乐，例如运动速度、运动显著性和时序（Di等人，2021）。我们的初步探索表明，这些特征在时间上密集分布，并持续影响音乐生成，导致速度不一致。CMT 使用 Lakh 钢琴卷帘数据集（Dong 等人，2018年），并采用扩展的复合词表示来处理音乐（Hsiao 等人，2021）。在这种表示中，每个词元都编码类型、节拍/小节标记、音符力度、音符密度、乐器、音高和时值。虽然这缩短了序列长度，却显著增加了输入维度。此外，CMT 中的音符事件严格与节拍细分对齐，而人类音乐家经常会引入表现性时序偏差——这是音乐演奏中传达情感的重要元素。在 CMT 的后续研究中，Zhuo 等人（2023年）依次训练了三个生成模型：一个用于生成和弦序列，另一个用于生成旋律，第三个用于生成伴奏。

Video2Music 模型同时利用低层视频特征和高层情感条件信息（Kang 等人，2024）。作者编制了 MuVi-Sync 数据集，其中包含 748 个带有音乐属性标签的音乐视频，例如音符密度、响度、和弦和调性，但不包含符号音乐数据。其编码器-解码器 Transformer 接收低层和高层视频特征，以及用户提供的引导和弦和调性，用于生成和弦序列。随后，使用固定模式将这些序列琶音化，从而生成最终的 MIDI 输出。然而，该模型依赖固定时间网格，因而无法保留人类演奏中细微的表现性节奏。此外，采用固定的琶音模式限制了音乐多样性，而必须由用户指定和弦与调性，也降低了非音乐家的使用便利性。

对基于视频的音乐生成器进行评估并非易事。在客观评价中，大多数研究会使用音级直方图熵、律动模式相似度和结构性指标（Wu 和 Yang，2020年）以及统计差异箱数量（Engel 等人，2019年）等指标，或对比损失（Alayrac 等人，2020年）等函数，衡量生成样本与真实值样本之间的差异。当数据集包含配对的视频和 MIDI 样本时，这种方法十分直接（Gan 等人，2020年；Zhuo 等人，2023年）。对于未配对的视频和 MIDI 样本，客观评估一个能够根据任意视频生成 MIDI 的模型并不可行。Su 等人（2021）使用各阶段对应训练数据集中的真实标注数据，评估其多阶段架构中的每个模块，具体包括节拍与风格预测、鼓轨创作，最终到音乐生成。

可控音乐 Transformer 的作者使用未配对数据集，评估在无条件生成、且没有输入视频的情况下得到的输出 MIDI（Di等人，2021）。他们通过实验表明，无条件生成的 MIDI 样本所呈现的指标更接近未配对 MIDI 数据集

与以视频为条件生成的输出样本相比。之所以出现这种情况，是因为用视频结构约束模型会导致偏离 MIDI 固有的结构，而要匹配某一特定 MIDI 数据集的指标，最好的方法就是仅在该数据集上进行训练。

总体而言，现有的基于视频的音乐生成方法存在若干局限。有些模型针对舞蹈或乐器演奏视频等狭窄场景设计，泛化能力有限。另一些方法只生成钢琴音乐，限制了音乐表达和乐器种类的范围。依赖连续时间条件控制的方法往往会扰乱自然的速度感，而几乎所有方法都使用固定且粗粒度的时间网格，无法捕捉人类音乐演奏中细腻的时序变化。

评估方法也存在明显不足。一些研究只评估多阶段模型的单个组件，而没有评估完整流程。另一些研究则孤立地评估 MIDI 生成，忽略其与视频内容的对齐。通过用户研究进行主观评估，仍是评估基于视频的 MIDI 生成最常见的实用方法（Di等人，2021；Gan 等人，2020年；Kang 等人，2024；Su 等人，2021；Zhuo 等人，2023年）。

我们继续讨论音频生成及其挑战，具体聚焦于音频带宽扩展任务。

2.7 音频带宽扩展

现代录音技术使音乐信号具有极高的音频质量。相比之下，爵士乐、流行音乐、民间音乐和布鲁斯等在20世纪60年代以前录制的档案录音，其聆听体验无疑受到当时录音技术以及物理介质老化的限制。此外，现代录音也可能因有损压缩、降采样、丢包或削波而导致音频质量下降。从最广义上说，音频增强旨在修复退化的信号，以改善其音质（Godsill 等，2002年）。因此，音频增强可以针对去除噪声、抑制爆裂声或噼啪声（例如来自老旧黑胶唱片的声音）、补全信号以填补缺失片段（即所谓的“音频修复”（Adler 等，2012年；Perraudin 等，2018年）），或扩展带限信号的带宽。

为了通过互联网流媒体传输音频信号，或便于存储，压缩、带宽缩减和低通滤波等常见操作都会导致至少部分高频音频内容被移除。虽然这一过程可以理解为将全带宽或宽带信号映射到带限或窄带信号，但与之对应的逆问题，即带宽扩展，旨在重建缺失的高频内容，因此并非易事。尽管如此，带宽扩展对于提高音频保真度至关重要，尤其是对语音和音乐信号而言。

接下来，我们将讨论音频带宽扩展中使用的数据集和方法。

2.7.1 数据集

音频带宽扩展模型以带限音频信号作为输入，并生成对应的全频带音频。带限音频是指频率成分受限于某个范围内的音频信号，通常低于完整的可听频谱（20 Hz - 20 kHz）。信号通过采样从模拟形式转换为数字形式，因此可表示的最高频率与数字信号的采样率直接相关。奈奎斯特定理指出，要准确重建模拟信号，所需的最低采样率是其最高频率成分的两倍。因此，在音频处理中，全频带信号通常以 44.1 或 48 kHz 的采样率进行采样。

为了研究音频带宽扩展任务，现有模型通常使用全频带（44.1+kHz）音频作为真值，并将其带限版本（即经过低通滤波的版本）用作输入。若干音乐源分离数据集包含全带宽音频，非常适合用于这项任务。尽管这些数据集提供了各个乐器的分离音轨，但全频带混音的可用性使其尤其适用于音频带宽扩展研究。DSD100（Demixing Secrets数据集）包含100首涵盖多种风格的音乐曲目，以及鼓、贝斯、人声和其他声部的分离音轨（Liutkus 等人，2017年）。类似地，MUSDB18 数据集包含150首来自多种类型的全频带曲目，每首曲目均提供鼓、贝斯、人声和其他乐器的分离音轨（Rafii 等人，2019年）。MedleyDB 数据集提供254首全频带歌曲及其各自的音轨，但其分离音轨采用了更细致、变化更多的标签分类，例如现场演奏中的观众噪声，或底鼓等单独的鼓组部件（Bittner 等人，2016）。

开放音频质量数据集（ODAQ）包含240个以44.1或48 kHz采样的音频样本（Torcoli等人，2024）。它包含由Netflix[10] 和Fraunhofer IIS[11]提供的流媒体音频。该数据集还包括每个样本的质量降级版本，旨在用于音频增强研究。ICASSP（国际声学、语音与信号处理会议）2023年声学回声消除（AEC）挑战赛引入了一个包含10,000个样本的全频带语音数据集，以及添加了合成回声效果的版本（Cutler等人，2023年）。合成复调环境声源（SPASS）数据集提供了25,000个采样率为44.1 kHz的音频样本。该数据集将短音频事件与较长的环境声音混合，旨在用于音频事件检测（Viveros-Muñoz等人，2023年）。

表2.4 总结了全频带音频数据集、其目标任务以及样本量 .

2.7.2 方法

音频带宽扩展最初的应用仅针对语音信号，这是由于电话系统带宽较低所带来的实际问题。最早的研究之一采用了一种统计方法，其中窄带和宽带频谱包络被假设为由窄带和宽带源的混合生成（Cheng 等，

[10]https://netflix.com[11]
https://iis.fraunhofer.de

表 2.4：全频带音频数据集

数据集	Task	样本量
MedleyDB（Bittner 等人，2016）	音乐源分离	254
Demixing Secrets 数据集（Liutkus 等人，2017年）	音乐源分离	100
MUSDB18（Rafii 等人，2017年）	音乐源分离	150
SPASS（Viveros-Muñoz 等人，2023年）	音频事件检测	25,000
AEC（Cutler 等，2023年）	声学回声消除	10,000
ODAQ（Torcoli 等，2024）	流式音频增强	240

1994年）。每个源对语音信号的贡献概率被参数化，然后使用期望最大化（EM）算法进行估计。基于映射的技术旨在学习窄带语音和宽带语音中特征之间的关联，其中的特征包括线性预测编码（LPC）系数和线谱频率（LSF）。基于码本映射的方法使用两个经过学习的码本，分别属于窄带信号和宽带信号，其中包含频谱包络特征，并且其条目之间存在一一映射关系（Epps & Holmes，1999年；Yoshida & Abe，1994年）。对于输入信号的每一帧，首先在窄带码本中找到最佳匹配条目，然后将其映射到宽带码本中对应的条目。基于线性映射的方法采用类似的方式，通过线性变换将窄带特征映射到宽带特征，其中变换矩阵使用最小二乘法等方法进行学习（Chennoukh 等，2001年；Nakatoh 等，1997年）。

后续方法尝试直接学习宽带信号的建模方式，而非预定义特征之间的映射。高斯混合模型（GMM）已用于估计窄带和宽带信号的联合概率密度（Nour-Eldin & Kabal，2008年；Park & Kim，2000b）。模型参数，即先验概率、均值向量和协方差矩阵，均由训练数据集学习得到。更复杂的方法包括使用隐马尔可夫模型（HMM），其中模型的每个状态都表示其窄带输入所对应的宽带扩展（Bauer & Fingscheidt，2008年；Jax & Vary，2003；Song & Martynovich，2009）。由于具有递归机制，HMM 可以利用过去输入帧的信息。基于非负矩阵分解（NMF）的方法将语音信号建模为学习得到的非负基的组合（Bansal 等，2005年；Sun & Mazumder，2013年）。在测试阶段，可以利用输入的低频基分量来估计如何组合高频基分量，以生成宽带信号。最后，早期将神经网络用于语音带宽扩展的研究，采用多层感知机（MLP）估计宽带语音信号的线性预测编码（LPC）系数（Iser & Schmidt，2003），或寻找用于变换频谱幅度的整形函数（Kontio 等，2007）。值得注意的是，这些早期研究使用的神经网络规模很小，参数总数约为 100。

较新的音频带宽扩展方法采用深度神经网络，其层数更多，表征能力也远胜于较早的方法。深度神经网络还消除了对

手工设计特征的需求，因为它们可以将原始音频或时频变换作为输入，然后学习适当的中间表示。最早采用深度神经网络进行音频带宽扩展的研究之一使用频谱图（K. Li 等人，2015年）。一项更深层的模型采用了流行的 *U-Net* 架构（Ronneberger 等人，2015年），并在原始音频领域中运行，对语音和单乐器音乐进行了实验（Kuleshov 等人，2017年）。T. Lim 等人（2018年）结合了上述两种方法，形成了一个双网络，分别在时域和频域中运行，并通过融合层生成最终输出。为了提升定性表现，即提高生成音频的清晰度，生成对抗网络（Goodfellow 等人，2020年）也被应用于基于深度神经网络的音频带宽扩展中（S. Kim 和 Sathe，2019年；Sautter 等人，2019年）。

老音乐录音的增强可以部分置于带宽扩展的背景下进行讨论，但在考虑提供给深度神经网络进行训练的数据时，会出现一些风险。尽管经过训练的深度神经网络在训练数据中的样本上表现良好，但在测试数据中的未见样本上，可能无法取得同样的表现。这种现象被称为样本过拟合，尽管这是一个重要的问题，尤其是在分类任务中，但它是否存在于图像超分辨率、音频带宽扩展和对抗生成等生成任务中，仍有争议。近期研究表明，在生成对抗网络的判别器和生成器中（Adlam 等人，2019年；Q. Xu 等人，2018年），以及用于视频帧生成的监督式生成网络中（Sulun，2018年；Sulun 和 Tekalp，2021），都未观察到样本过拟合。此外，最先进的图像超分辨率神经网络不包含批归一化（Ioffe 和 Szegedy，2015年）和随机失活（Srivastava 等人，2014年）等正则化层，以避免过拟合（Fan 等人，2017年；B. Lim 等人，2017年；Y. Zhang 等人，2018年）。

2.8 后续章节概述

至此，我们完成了对背景的讨论，并将转而介绍我们的创新贡献和总体方法。我们首先开发一个独立的条件音乐生成器，不依赖视频输入。最初，模型以情感和时间信号为条件。然而，基于情绪的音乐生成需要带有情绪标签的符号音乐数据。正如第 2.4.3.1 节所述，现有的情感标注 MIDI 数据集规模过小，无法有效训练深度神经网络。因此，在第 3 章中，我们将通过创建一个大规模情感标注 MIDI 数据集来解决这一局限。

有了这个数据集后，我们将在第4章聚焦于条件音乐生成。在建立了独立的音乐生成器后，我们将在第5章转而关注视频分析。在第6章，我们将介绍如何将条件音乐生成器与视频分析模块整合。我们将在第7章结束方法论部分的讨论，探讨音频生成面临的挑战，尤其是音频带宽扩展。最后，在第8章，我们将总结全文并概述未来的研究方向。

第 3 章

MIDI 情绪标注

条件深度神经网络经过训练后，可根据特定输入条件生成输出。训练此类模型需要一个同时包含目标输出及其对应条件的数据集。训练过程中，条件作为输入提供给深度神经网络，随后由其生成输出。模型通过最小化损失函数进行优化，该函数用于衡量生成输出与真实目标之间的差异。推理时，只需提供条件，训练好的深度神经网络便能生成与给定条件相对应的输出。

基于视频的音乐生成器的核心组件之一是条件音乐生成器。尽管在我们研究的后续阶段中，条件化输入最终将由视频推导得出，但我们首先通过手动提供条件来设计和测试音乐生成器。

我们的最终目标是利用高层次情感特征和低层次时间特征，建立输入视频与输出 MIDI 之间的对应关系。低层次时间特征可以直接从 MIDI 数据中提取，因此这一阶段不需要任何额外的标签。相比之下，MIDI 文件不包含高层次情感信息。因此，要训练情感条件 MIDI 生成器，我们需要带有情绪标签的 MIDI 数据集。我们将在第 4 章中介绍这些标签如何用于 MIDI 生成。

由于我们的目标是训练深度神经网络以改进音乐生成，因此需要大型数据集。然而，正如第2.4.3.1 节所述，现有的带情感标签的 MIDI 数据集规模过小，无法支持此类训练。为此，我们为大规模纯 MIDI 数据集提供情感标注。鉴于这些数据集规模庞大，手动标注并不切实际。因此，我们提出一种利用音频和文本等其他模态的自动标注方法。具体而言，我们首先利用歌曲的 MIDI 和音频版本之间的对齐关系。然后，我们探究歌曲的 MIDI 表征与其歌词之间潜在的情感关联。

Spotify[1] 为其曲库中的歌曲同时提供低层特征和高层特征。我们的分析表明，特征可以有效地表征音乐情感。此外，由于

[1] https://spotify.com

部分MIDI 文件包含嵌入的歌词，我们还可以利用文本处理模型从歌词内容中提取情感信息。接下来，我们将详细介绍这两种方法。

3.1 使用 Spotify 特征

Spotify 开发者 API（应用程序编程接口）[2] 可自动访问歌曲的音频特征。通过此 API，我们获取了多种音乐属性，包括舞蹈性、能量、调性、响度、调式、言语性、原声度、器乐性、现场感、效价和速度。其中，效价对于表示音乐情感尤其重要。尽管计算这些特征所采用的确切方法并未公开，但官方文档提供了详细说明[3]。

利用此 API，我们的目标是为现有最大的 MIDI 文件合集——Lakh MIDI 数据集（LMD）标注情绪。为此，我们必须先在 Spotify 曲库中找到对应条目。幸运的是，LMD 包含一个名为*LMD-matched*的子集，其中一部分 MIDI 文件与百万歌曲数据集（MSD）的条目相匹配（Bertin-Mahieux 等人，2011年）。每个 MSD 条目都提供元数据，例如歌曲标题、艺术家名称，以及指向另一个音乐数据库 The Echo Nest 的标识符[4]。百万歌曲数据集 Echo Nest 映射档案[5] 通过提供对应于 Echo Nest 的 Spotify 曲目 ID，进一步完善了这一关联。总之，我们沿着这条映射链，从 LMD-matched 出发，经过 MSD 和 Echo Nest，最终将 MIDI 文件与对应的 Spotify 曲目关联起来。

然而，MSD 中并非所有条目都关联了 Spotify 曲目 ID。为扩大带标签数据集，我们采用另一种策略：利用 MSD 提供的艺术家姓名和歌曲标题，直接在 Spotify 数据库中进行搜索。

Spotify 音频属性中包含 效价 特征，使我们能够将其直接整合到情感环形模型中（第 2.2.2 节，图 2.4）（Russell，1980），该模型使用效价和唤醒度来表示 MIDI 曲目的情感。这使我们的任务简化为对唤醒度特征进行建模。然而，我们对 Spotify 特征的初步分析表明，能量特征并不适合用于唤醒度建模。我们推测，这种差异源于音频与符号音乐之间的音色差异，这意味着同一首作品会因录音的不同而呈现出不同的能量水平（例如音量、音调、音效）。

为对唤醒度维度进行建模，我们借鉴现有文献，采用可直接从 MIDI 中提取的低层级特征（H. H. Tan 和 Herremans，2020年；Williams、Kirke、Miranda 等，2015年）。我们的实验表明，速度和音符密度在表示唤醒度方面尤为有效。

我们使用 MIDI 文件内置的元数据来估算速度。刻度分辨率表示 MIDI 文件的时间分辨率。例如，刻度分辨率为 0.001，意味着 MIDI 事件

[2] https://developer.spotify.com[3] https://developer.spotify.com/documentation/web-api/reference[4] https://en.wikipedia.org/wiki/The_Echo_Nest[5] https://labs.acousticbrainz.org/million-song-dataset-echonest-archive

之间至少可以相隔 1 毫秒。分辨率指每个音乐节拍包含的时钟刻度数。在音乐中，节拍是时间的基本单位，通常表现为一种节奏脉冲，用来确定一首乐曲的速度和结构。将刻度分辨率乘以分辨率，即可得到一个节拍的时值（秒）。要估算每分钟节拍数（BPM），我们用一分钟的时长除以一个节拍的时值。以下公式总结了这一方法：

$$\text{Estimated tempo} = \frac{60.0}{\text{Tick scale} \times \text{Resolution}} \tag{3.1}$$

我们人工审听了一组样本，凭耳朵估算其速度，并将这些估算值与计算结果进行比较。我们发现，当 MIDI文件包含鼓轨时，估算结果与实际情况较为吻合；可以说，鼓轨是确立节奏时影响最大的乐器。然而，在没有鼓轨时，我们的方法偶尔会失效，尤其是在仅钢琴曲目中。在后续基于情感的音乐生成研究中，我们仅在存在鼓轨时尝试将估算速度用作唤醒度。

我们还提取音符密度，其定义为所有乐器每秒演奏的音符平均数量。此外，我们还计算乐器数量，以及平均音符密度，后者通过音符密度除以乐器数量得出。即使没有鼓轨，将平均音符密度用作唤醒度特征也有效。不过，对条件音乐生成输出进行人工分析后发现，与使用估算速度相比，使用平均音符密度作为唤醒度的可靠性较低。

我们的整个数据集创建流程如图所示3.1。

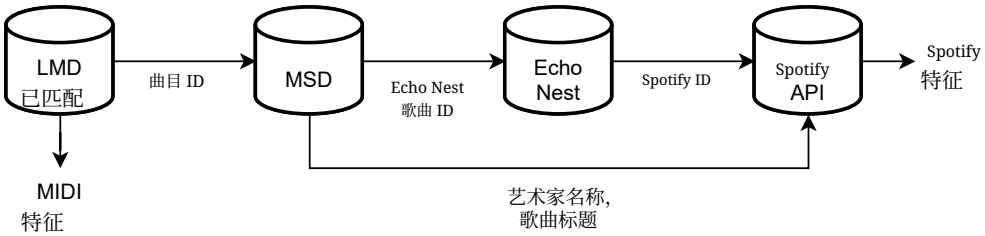


图 3.1：数据集创建流程。

为了保证完整性并促进未来研究，我们为整个 Lakh MIDI 数据集提取了低层 MIDI 特征，并将其标记为 *LMD-full*，其中并非所有样本都必然映射到 MSD 中的条目。LMD-full 数据集包含 178,561 个 MIDI 文件。经过调查，我们发现其中 174,270 个有效，并丢弃了其余损坏或空白的文件。Lakh MIDI 数据集是通过从互联网上公开可用的来源下载 MIDI 文件，再根据其哈希值保留唯一文件而构建的。然而，经进一步检查，我们发现哈希值不同的 MIDI 文件仍可能具有相同的音乐内容，这可能是由于元数据存在差异。为了进一步过滤数据并仅保留具有独特音乐内容的 MIDI 文件，我们首先将 MIDI 文件转换为钢琴卷帘，使用

pretty_midi 包[6]。然后，我们按字母顺序拼接所有乐器的二维钢琴卷帘矩阵，从而创建一个仅描述音乐内容的单一二维矩阵。在使用这些矩阵重新计算哈希值后，我们选取哈希值唯一的文件。该过程得到了 152,968 个具有独特音乐内容的 MIDI 文件。这些就是我们最终用于训练音乐生成模型的文件。

Lakh MIDI 数据集的匹配划分 LMD-matched 包含 31,034 首来自 MSD 的曲目，这些曲目与 LMD 中的 116,189 个 MIDI 文件相匹配。由于多个 MIDI 文件可能匹配同一首曲目，多个曲目也可能匹配同一个 MIDI 文件，因此我们采用了滤波处理，仅保留音乐内容唯一的 MIDI 文件，做法与 LMD-full 划分相同。此外，我们根据 LMD-matched 子集中提供的匹配分数，为每个 MIDI 文件仅保留 MSD 中匹配度最高的曲目，以确保每个 MIDI 文件只有一组标签。筛选出内容唯一且有效的 MIDI 文件后，我们获得了 36,545 个与 MSD 条目相匹配的 MIDI 文件。利用 MSD 中的元数据，我们查询了 Spotify 的数据集，并成功检索到其中 34,791 个 MIDI 文件的音频特征。

在完整形式下，我们的数据集包含舞蹈性、能量、调性、响度、调式、言语性、原声度、器乐性、现场感和效价等高层标签，以及音符密度、速度和乐器数量等低层 MIDI 特征。表 3.1 展示了我们的数据集与现有带有情绪标签的 MIDI 数据集之间的比较。该数据集构成了基于情感的 MIDI 生成器的训练数据集，相关内容将在第 4.2 节讨论。清单 3.1 展示了一个样本条目及其包含的特征。

表 3.1：带情感标签的 MIDI 数据集

数据集	歌曲数量 歌曲	时值 （小时）	配器	情绪标签
类似 MIREX （Panda 等人，2013年）	193	11.6	多乐器	类别型
VGMIDI (Ferreira& Whitehead, 2019年)	204	6.4	仅钢琴	维度
EMOPIA (Hung 等人，2021)	387	11.0	仅钢琴	类别型
Ours	34791	2161.9	多乐器	维度

[6]https://craffel.github.io/pretty-midi

清单 3.1：所提数据集中的一个样本条目。

```
"cc992d0d8e82d09b7fe2466cf851497a":{

"我的标识符_功能":{
"无te_de密度": 30.364985431879415,
"速度": 84.000084000084,
"n_乐器": 10
},
"匹配的_特征": {
"曲目_id": "TRUHHPK12903CBA84F" ,
"匹配 _分数": 0.7362919446232072 ,
"歌曲_ID": "SOSYWZT12AB0187E45",
"标题": "In 《夏日时光》" ,
"艺术家": "Mungo Jerry",
"发行": "Uber 30 – dasrockt!",
"spotify_id":"5VPOrzHyuULaiCKnwQNNCN",
"spotify_title": "In The Summertime",
"Spotify_艺人": "Mungo Jerry",
"Spotify_专辑": "Uber 30 – 摇滚乐!",
"spotify_audio_音频特征":{
"可舞性": 0.681,
"能量": 0.509,
"调性": 4,
"响度": -8.504 ,
"调式": 1,
"语音感": 0.0461,
"声学性": 0.497 ,
“工具性”: 2.72e-06,
“活跃度”: 0.188 ,
"valence": 0.963 ,
"速度": 82.614,
"类型": "音频_特征",
"id":"5VPOrzHyuULaiCKnwQNNCN",
"uri":
"spotify:track:5VPOrzHyuULaiCKnwQNNCN",
"trackhref": "https://api.spotify.com/v1/"
tracks/5VPOrzHyuULaiCKnwQNNCN",
"分析_网址":"https://api.spotify.com/v1/
音频-分析/5VPOrzHyuULaiCKnwQNNCN",
"时长 _毫秒": 210387 ,
"时间_签名": 4
}
}
}
```

3.2 歌曲歌词的情绪分类

我们获取 MIDI 样本情绪标签的第二种方法，是分析 MIDI 文件中所包含歌曲歌词的情绪。MIDI 文件主要用于存储音符事件，但也可以选择包含曲目名称、版权声明和歌词等文本元数据。

我们的方法利用了歌词与音乐之间天然的联系，两者往往都与情绪相关（Stratton & Zalanowski，1994 ）。与 Spotify 音乐特征类似，我们注意到，歌词中的情绪内容也能为 MIDI 样本提供弱特征，尽管有时歌词与音乐所表达的情绪会相互冲突。一个著名的例子是 The Police 的歌曲《*Every Breath You Take*》：音乐传达出充满爱意与关怀的氛围，而歌词则以跟踪者的视角写成，反映了不健康的执念、嫉妒和控制欲[7]。

Spotify 开发者 API 会直接提供效价值，而歌词中的情绪特征则必须通过文本分析提取。我们的方法是使用文本情绪分类器分析歌曲歌词，并将预测出的情绪分配给 MIDI 样本。为此，我们首先在 GoEmotions 数据集上训练情绪分类模型。该数据集是规模最大的文本数据集之一，包含 28 种细粒度情绪标签（Demszky 等人，2020年）。

3.2.1 训练

实现构建情感标注符号音乐数据集这一目标的第一步，是训练一个模型，使其能够根据文本输入进行多标签情绪分类。

3.2.1.1 数据集

我们使用 GoEmotions 数据集（Demszky 等人，2020年）训练模型。该数据集由 Reddit 上的英文评论构成[8], ，这些评论经过人工标注，以识别其中隐含的情绪。这是一个多标签数据集，这意味着每个样本可以包含多个情绪标签。该数据集包含 27 种情绪和一个“中性”标签。Demszky 等人（2020年）进一步将这些标签归为 7 个类别，其中包括 Ekman （ 1971 年）提出的六种基本情绪（喜悦、愤怒、恐惧、悲伤、厌恶和惊讶）以及“中性”标签，如表3.2 所示。该数据集共包含 5.8 万个样本，并分别按 80%、10% 和 10% 的比例划分为训练集、验证集和测试集。鉴于其标签数量和规模，GoEmotions 是规模最大的情绪分类数据集之一，也是离散情绪标签数量最多的数据集（Kusal 等人，2022年）。

[7] https://www.bbc.co.uk/radio2/soldonsong/songlibrary/indepth/everybreathyoutake.shtml
[8] https://www.reddit.com

表 3.2：GoEmotions 标签与 Ekman类别的映射

GoEmotions 标签	**Ekman类别**
愤怒、恼怒、不赞同	愤怒
厌恶	厌恶
fear, 紧张	fear
钦佩、愉悦、认可、关怀、渴望，兴奋、感激、喜悦、爱、乐观、自豪、宽慰	joy
悲伤、失望、尴尬、悲痛、懊悔	悲伤
困惑、好奇、恍然大悟、惊讶	惊讶
中性	中性

3.2.1.2 实现细节

我们采用 DistilBERT 作为模型的骨干网络（Sanh 等人，2019年），它是通过知识蒸馏（Bucila 等人，2006；Hinton 等人，2015年）从 BERT（来自 Transformer 的双向编码器表示）模型（Devlin 等人，2019年）精简压缩而成的变体。DistilBERT 使用的层数少于 BERT，并通过学习 BERT 的输出来模仿其行为。我们的模型由 6 层组成，每层包含 12 个注意力头，维度为 768，总计拥有 6700 万个参数。为实现多标签分类，我们对输出层进行了定制，并在末尾添加 Sigmoid 激活层，而非使用适用于单标签分类的 Softmax 层。输出层的大小由训练数据集中标签的数量决定，可以是 7 或 28。

我们使用二元交叉熵损失训练模型。在评估中，我们采用精确率、召回率和F1分数，并进行宏平均。我们搜索最优决策阈值，以最大化F1分数。决策阈值设置为0.3，这意味着值为0.3或更高的预测被视为正类预测，其余则视为负类预测。

我们训练两个模型，将给定文本分别分类为7个和28个标签。我们采用0.1的丢弃率和1的梯度裁剪范数。对于具有7个输出标签的模型，批量大小设置为16；对于具有28个输出标签的模型，批量大小设置为32。前者采用5e-5的学习率，后者采用3e-5的学习率。我们根据验证数据集上的F1分数使用提前停止机制，这对应于两个模型均训练10个轮次。我们使用Huggingface库（Wolf等人，2019年）和PyTorch后端（Paszke等人，2019年）实现模型，并使用单个Nvidia GeForce GTX 1080 Ti GPU对其进行训练。

3.2.2 推理

在对基于文本的情感分类模型进行训练后，我们在推理模式下使用该模型，将来自 MIDI 文件的歌曲歌词作为输入。这使我们能够创建一个标注了情绪的 MIDI 数据集。类似于在第 3.1 节中使用音频相关标签，我们将这些视为“弱标签”

因为它们源自文本而非 MIDI，并且由经过训练的模型生成，而非由人类标注。

我们使用两个公开的 MIDI 数据集，这些数据集由汇集各种在线来源的 MIDI 文件构建而成：包含 176k 个样本的 Lakh MIDI 数据集（Raffel，2016）以及包含 130k 个样本的 Reddit MIDI 数据集[9]。我们通过筛选包含至少 50 个英语单词歌词的 MIDI 文件，对这些数据集进行筛选。筛选后共得到 12509 个文件，其中 8386 个来自 Lakh MIDI 数据集，4123 个来自 Reddit MIDI 数据集。在推理过程中，我们使用了两个预训练模型，并输入整首歌曲的歌词，截断长度设为 512 个词元。

3.2.3 结果

本节首先介绍训练模型的情绪分类性能。随后，我们介绍情感标注 MIDI 数据集；该数据集是我们使用训练好的模型分析歌曲歌词的情感后创建的。

3.2.3.1 GoEmotions 数据集上的情绪分类

我们在 GoEmotions 数据集的测试集上评估了训练模型的演奏效果，并将结果与原论文（Demszky 等人，2020年）中给出的基线进行了比较。与原论文类似，我们报告了使用两组标签的场景结果，分别包含 7 种和 28 种情绪。对于每个标签，我们报告精确率、召回率和 F1分数，以及宏平均值。需要指出的是，由于数据集存在类别不平衡，宏平均比微平均更为适用，原论文也采用了宏平均。我们注意到，基线模型为 BERT，其规模是我们模型的两倍（Devlin 等人，2019年）。

精确率与召回率之间的权衡由阈值决定。因此，我们更强调较高的 F1分数，因为 F1分数通过计算精确率和召回率的调和平均值，能够提供更均衡的视角，并且对阈值不那么敏感。尽管原论文未说明阈值，但在 7标签数据集上使用 0.3 作为阈值时，我们取得了最佳 F1分数，并获得了与原论文相近的演奏效果。为保持一致，我们对 28标签数据集使用相同的值。我们在表 3.3 和 3.4 中分别呈现了包含 7 个和 28 个标签的数据集上的结果。值越高表示演奏效果越好，最佳结果以粗体突出显示。

根据 F1分数，我们的模型在 7标签数据集上的表现与基线相当。具体而言，我们的模型在 2 个标签上的表现更好，在 2 个标签上更差，在 3 个标签上持平，宏平均也是如此。在 28标签数据集上，我们的模型优于基线，仅在 2 个标签上的表现较差，在 4 个标签上持平，在其余 22 个标签上表现更好。此外，我们的模型在宏平均方面提高了 0。04。

[9]https://archive.org/details/themagicofmidiv1

表 3.3：7标签分类结果

	精确率		召回率		F1分数	
	基线	Ours	基线	Ours	基线	Ours
愤怒	0.50	0.50	0.65	0.67	**0.57**	**0.57**
厌恶	0.52	0.57	0.53	0.49	**0.53**	0.52
fear	0.61	0.57	0.76	0.73	**0.68**	0.64
joy	0.77	0.75	0.88	0.89	**0.82**	**0.82**
中性	0.66	0.63	0.67	0.75	0.66	**0.68**
悲伤	0.56	0.57	0.62	0.67	0.59	**0.61**
惊讶	0.53	0.59	0.70	0.62	**0.61**	**0.61**
宏平均	0.59	0.60	0.69	0.69	**0.64**	**0.64**

表3.4：28标签分类结果

	精确率		召回率		F1分数	
	基线	Ours	基线	Ours	基线	Ours
钦佩	0.53	0.65	0.83	0.75	0.65	**0.70**
愉悦	0.70	0.72	0.94	0.91	0.80	**0.81**
愤怒	0.36	0.53	0.66	0.49	0.47	**0.51**
恼怒	0.24	0.40	0.63	0.31	0.34	**0.35**
认可	0.26	0.39	0.57	0.38	0.36	**0.39**
关怀	0.30	0.37	0.56	0.46	0.39	**0.41**
困惑	0.24	0.52	0.76	0.42	0.37	**0.47**
好奇	0.40	0.47	0.84	0.62	**0.54**	0.53
渴望	0.43	0.66	0.59	0.42	0.49	**0.51**
失望	0.19	0.39	0.52	0.22	**0.28**	**0.28**
不赞同	0.29	0.39	0.61	0.41	0.39	**0.40**
厌恶	0.34	0.64	0.66	0.39	0.45	**0.48**
尴尬	0.39	0.72	0.49	0.35	0.43	**0.47**
兴奋	0.26	0.43	0.52	0.47	0.34	**0.45**
fear	0.46	0.60	0.85	0.76	0.60	**0.67**
感激	0.79	0.88	0.95	0.92	0.86	**0.90**
悲痛	0.00	0.00	0.00	0.00	**0.00**	**0.00**
joy	0.39	0.59	0.73	0.61	0.51	**0.60**
love	0.68	0.78	0.92	0.85	0.78	**0.81**
紧张	0.28	0.45	0.48	0.43	0.35	**0.44**
中性	0.56	0.61	0.84	0.76	**0.68**	**0.68**
乐观	0.41	0.56	0.69	0.52	0.51	**0.54**
自豪	0.67	0.83	0.25	0.31	0.36	**0.45**
恍然大悟	0.16	0.39	0.29	0.14	**0.21**	**0.21**
宽慰	0.50	0.00	0.09	0.00	**0.15**	0.00
懊悔	0.53	0.59	0.88	0.86	0.66	**0.70**
悲伤	0.38	0.57	0.71	0.60	0.49	**0.59**
惊讶	0.40	0.56	0.66	0.50	0.50	**0.53**
宏平均	0.40	0.53	0.63	0.50	0.46	**0.50**

我们假设，在某些情况下，较小的模型（如我们的方法 DistilBERT）可能比更大的基线模型（BERT）表现更好，例如训练样本数量有限，或输出/目标维度较高时，28标签数据集便是这种情况。在这些情形下，模型更容易过拟合，这一点此前已有观察（Yu 等，2017年）。此外，原论文（Sanh 等，2019年）表明，DistilBERT 模型在 Winograd 自然语言推理（WNLI）数据集上的表现优于 BERT（Levesque 等，2012年）。

3.2.3.2 MIDI 数据集上的推理

我们使用训练好的模型分析 Lakh 和 Reddit MIDI 数据集中的歌曲歌词，从而得到一个增强数据集，其中包含 12509 个 MIDI 文件的文件路径及其对应的情绪标签预测概率。为给用户提供更大的灵活性，我们不对预测概率应用阈值，而是提供原始分类输出，使整个数据集可以直接使用。我们生成两个 CSV（逗号分隔值）文件，分别以 7 个和 28 个情绪标签作为列，以 12509 个 MIDI 文件路径作为行。

为便于演示，我们提供了表格的转置版本，仅使用 3 个样本，如表 3.5 和 3.6所示。为进一步演示并便于分析，我们在清单 3.2中提供了三首样本歌曲的歌词摘录，并按降序列出预测概率高于 0.1 的情绪。值得注意的是，包含 28 个情绪标签的数据集能够更细致地表征情绪。例如，在查看该数据集时，我们发现歌曲《Imagine》最有可能被预测为具有“乐观”情绪，而《Take a Chance on Me》最有可能被预测为具有“关怀”情绪。不过，在仅包含七个标签的数据集中，这两首歌最有可能被预测为具有“喜悦”情绪。

我们还在图 3.2 中展示了数据集中包含每种情绪的样本数量。在这些图中，我们排除了“中性”标签。为便于演示，我们还将预测值高于 0.1 的情绪视为正向标签，即这些情绪存在于相应样本中。由于篇幅有限，文件路径替换为艺术家和歌曲名称，具体如下：约翰·列侬 - Imagine：“`lakh/5/58c076b72d5115486c09a7d9e6df1029.mid`”（艺术家和片名信息取自百万歌曲数据集（Bertin- Mahieux 等，2011）），ABBA - Take a Chance on Me：“`reddit/A/ABBA.TakeachanceonmeK.mid`”，埃尔维斯·普雷斯利 - Are You Lonesome Tonight：“`reddit/P/PRESLEY.AreyoulonesometonightK.mid`”

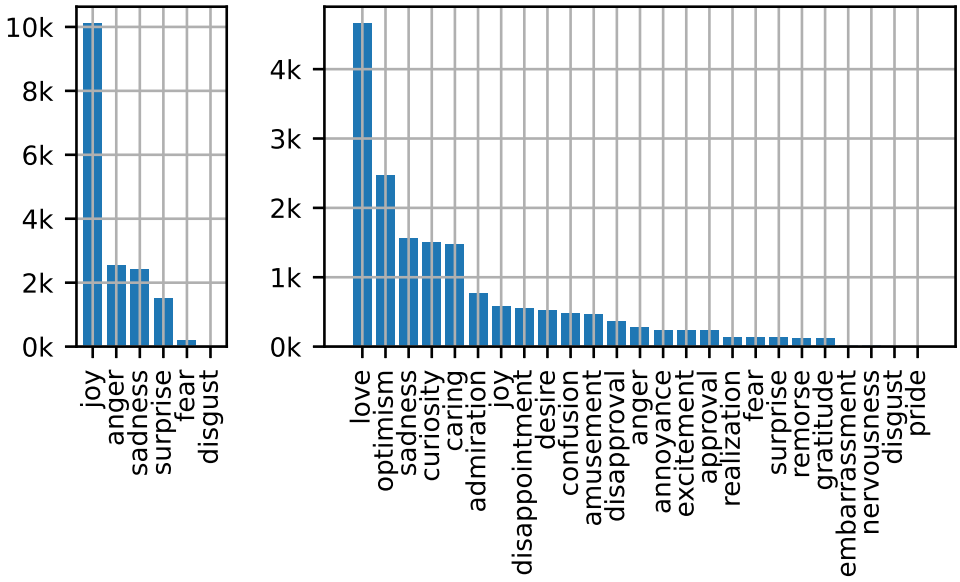


图3.2：我们的7标签数据集（左）和28标签数据集（右）中包含每种情感的样本数量。

表3.5：7标签数据集中的样本条目。

	约翰·列侬 - Imagine	ABBA - Take 给我一个机会	埃尔维斯·普雷斯利 - Are You Lonesome Tonight
愤怒	0.0051	0.0146	0.0272
厌恶	0.0003	0.0009	0.0045
fear	0.0005	0.0024	0.0131
joy	**0.8072**	**0.8948**	0.0477
中性	0.1953	0.1420	0.0782
悲伤	0.0013	0.0069	**0.7372**
惊讶	0.0754	0.0053	0.5465

表 3.6：28标签数据集中的样本条目。

	约翰·列侬 - Imagine	ABBA - Take a Chance on Me	埃尔维斯·普雷斯利 - Are You Lonesome Tonight
钦佩	0.0021	0.0091	0.0048
愉悦	0.0051	0.0012	0.0027
愤怒	0.0025	0.0018	0.0053
恼怒	0.0024	0.0020	0.0075
认可	0.0026	0.0809	0.0072
关怀	0.0067	**0.6169**	0.0601
困惑	0.0070	0.0035	0.1029
好奇	0.0332	0.0141	**0.6502**
渴望	0.0482	0.0472	0.0055
失望	0.0044	0.0016	0.0199
不赞同	0.0019	0.0030	0.0048
厌恶	0.0007	0.0003	0.0009
尴尬	0.0006	0.0002	0.0045
兴奋	0.0130	0.0049	0.0011
fear	0.0026	0.0026	0.0035
感激	0.0007	0.0017	0.0059
悲痛	0.0008	0.0016	0.0085
joy	0.0025	0.0040	0.0018
love	0.0021	0.1079	0.0193
紧张	0.0007	0.0017	0.0094
中性	0.2954	0.4288	0.0757
乐观	**0.7554**	0.1423	0.0060
自豪	0.0010	0.0013	0.0006
恍然大悟	0.0023	0.0040	0.0045
宽慰	0.0004	0.0033	0.0011
懊悔	0.0005	0.0012	0.1491
悲伤	0.0011	0.0027	0.1767
惊讶	0.0107	0.0005	0.0020

列表3.2：歌词摘录样本，以及预测值高于 0.1 的情绪。

文件路径：lakh/5/58c076b72d5115486c09a7d9e6df1029.mid 艺术家 - 片名：John Lennon - Imagine 歌词：想象一下，那里没有天堂。这很容易，只要你尝试。我们脚下没有地狱。我们头顶只有天空。想象所有的人们。为今天而活。7 类标签预测：喜悦：0.8072 中性：0.1953 28 类标签预测：乐观：0.7554 中性：0.2954

文件路径：reddit/A/ABBA.Take a chance on me K.mid 艺术家 - 片名：ABBA - Take a Chance on Me 歌词：如果你改变主意，我会第一个排队。亲爱的，我依然自由。给我一个机会。如果你需要我，请告诉我，我会一直在这里。如果你无处可去，如果你感到沮丧。7 类标签预测：喜悦：0.8948 中性：0.1420 28 类标签预测：关怀：0.6169 中性：0.4288 乐观：0.1423 爱：0.1079

文件路径：reddit/P/PRESLEY.Are you lonesome tonight K.mid 艺术家 - 片名：埃尔维斯·普雷斯利 - Are You Lonesome Tonight 歌词：Are you lonesometonight? Do you miss me tonight? Are you sorry wedriftedapart? Does your memorystray to a bright summer day, When I kissed you and called you sweetheart? 7 类标签预测：悲伤：0.7372 惊讶：0.5465 28 类标签预测：好奇：0.6502 悲伤：0.1767 懊悔：0.1491 困惑：0.1029

3.3 讨论

在对符号音乐自动情绪标注的探索中，我们研究了两种主要方法：利用 Spotify 的音频相关特征，以及分析嵌入 MIDI 文件中的文本歌曲歌词。我们的目标是为大规模 MIDI 数据集补充情绪元数据，以便训练条件音乐生成器。初步评估结果显示，从 Spotify 特征推断出的情绪标签比从歌词推断出的标签更可靠。因此，我们最终采用了基于 Spotify 得出的标签来进行情绪条件 MIDI 生成。

Spotify 提供了一系列高层次的音乐特征，其中效价尤其与情感环形模型等成熟的情绪模型高度吻合。我们利用 Lakh MIDI 数据集与 Spotify 曲库之间的关联，通过百万歌曲数据集和 Echo Nest ID 建立匹配。这个过程为数以万计的 MIDI 文件生成了情绪元数据，为基于情绪的 MIDI 生成奠定了切实可行的基础。

我们也尝试使用歌词推断情绪内容，但这种方法在我们的场景中效果较差。尽管我们使用 GoEmotions 数据集和 DistilBERT 主干网络训练了一个多标签情感分类器，但根据歌词生成的标签缺乏一致性，无法在下游生成任务中产生感知上连贯的结果。这一局限可能源于歌词的歧义、歌词情绪与音乐情绪之间的不匹配，以及互联网文本与歌曲歌词之间的领域差距。因此，我们未将基于歌词的标注纳入最终生成流程。

我们在基于文本的情感分类方面的定量结果凸显了 DistilBERT 在情绪分类上的高效性，尤其是在输出标签数量较多的情况下，例如 28 种情绪标签集。值得注意的是，我们的模型在 28标签数据集上优于基线，这表明较小的模型能够避免过拟合，并且在训练数据有限、类别众多的情况下仍能取得出色的结果。将文本分为 28 个细粒度情绪类别的能力，使我们能够捕捉歌曲中更细腻的情绪变化，如清单 3.2 中的示例输出所示。

尽管我们在音乐生成流程中排除了基于歌词的标签，但根据歌词进行情感分类仍是一个很有前景的研究方向。歌词能够传达歌曲创作者的情感意图，并在音频特征所揭示的信息之外提供一个语义层面。通过改进特定领域的模型，或采用将歌词与音频及符号数据相结合的多模态策略，可以实现更细致、更准确的情感标注。我们提供的数据集和模型旨在支持这些研究方向，尤其是基于情绪的音乐检索或情感感知歌词创作等任务。

综上，这些标签方法为情感条件音乐生成奠定了基础，使符号音乐能够根据目标情感特征进行创作或引导。下一章将介绍如何利用这些标签训练能够响应情感条件的生成模型。

第4章

条件音乐生成

本章将介绍我们的条件音乐生成方法。该方法是基于视频的音乐生成系统中最重要的组成部分，因为该模块负责生成最终的音乐输出。虽然本章中的输入条件由人工提供，但在第6 章中，我们将用通过视频分析获得的结果替代这些条件。正如前文所述，我们选择先研究条件音乐生成，再在第5章研究视频分析，以确保研究的每个阶段都能产生具有音乐意义的结果。

如第1章所述，我们的目标是实现输入视频与生成音乐之间高层和低层的对齐。我们将情绪作为高层特征，将时间边界作为低层中间特征。这样，生成的配乐既能反映视频的情感基调，又能保持时间同步。

我们首先概述时间条件音乐生成的方法。

4.1 基于时间边界的音乐生成

视频与音乐的时间对齐对于营造“同步”的感觉至关重要（Prender- gast，1992）。我们的方法旨在克服最先进方法所采用的密集时间条件机制的局限性（Di等人，2021），相关讨论见第1.1.1节。为此，我们提出使用稀疏时间特征，具体而言即时间边界。在这种方法中，我们首先定义视频和音乐中的时间边界。

由于场景（或等价地说，镜头）是视频的基本单位，我们使用场景切换来标记时间边界（Cutting，2005年；Nitanda等人，2005年）。在音乐领域，我们采用和弦位置作为时间边界，因为和弦经常在音乐作品中充当结构锚点（Bharucha 和 Krumhansl，1983）。通过对齐这两种模态，我们旨在训练一个音乐生成器，使其在输入视频的场景边界附近安排和弦，同时保持生成旋律整体的音乐连贯性。

为实现时间条件音乐生成，我们首先提取和弦位置，作为训练期间的输入。我们在训练数据集——Lakh Pianoroll Dataset-5（Dong 等人，2018年）中标注和弦。该数据集是 Lakh MIDI 数据集（Raffel，2016）的一个版本，其中乐器被合并为五类

：贝斯、鼓、吉他、钢琴和弦乐。我们重点关注吉他和钢琴中至少包含三个同时发音的音符，且持续时间至少为两拍的和弦。我们将其称为强和弦，并指出这一术语与其在和声语境中的任何潜在用法有所不同。为标注这些强和弦，我们通过添加CHORD 词元来扩充标记词汇表。每个强和弦都通过在其第一个ON （即NOTE-ON）词元之前插入一个CHORD 词元来标记。为了使模型不仅能在视频场景切换位置生成和弦，也能在音乐上合适的位置独立生成和弦，我们在训练期间随机移除 20% 的CHORD 词元。

和弦是旋律不可或缺的一部分，为前后相邻的音符提供和声与节奏支撑（Bharucha 和 Krumhansl，1983）。在推理过程中，若在序列的特定位置强制插入CHORD 词元，可能导致和弦听起来不合拍或过于突兀。这是因为模型事先并不知道即将出现的和弦，因而可能生成与之无法自然衔接的前置音符。为解决这一问题，我们提出一种方法，使模型能够“预判”即将出现的和弦，并据此生成前置音符与时间偏移。此外，我们训练模型自行生成CHORD 词元，以确保生成音乐的节奏一致性。为此，我们为每个输入标记定义边界偏移量，表示距离下一个边界的剩余时间。这些偏移量设有上限；并且由于我们的音乐生成器是自回归的，它们是根据未来和弦而非过去和弦计算的。此外，我们将CHORD词元的损失放大 10 倍，因为它会影响多个前置和弦音符，并在构建生成音乐的结构中起着关键作用。

在图 4.1中，我们阐释了相对于强和弦的边界偏移量这一概念。图的上半部分展示了符号音乐，其中包含三个或更多同时发音的音符且持续时间超过特定阈值的和弦定义了一个音乐边界。图的下半部分展示了边界偏移量，表示到下一个边界的时间距离。需要注意的是，这些图仅用于说明；除了定义的边界处，偏移量并不会与音乐完全对齐。

我们使用前馈网络处理边界偏移量，生成边界偏移编码，然后将其与学习到的位置编码（定义见第 2.3节）（B.

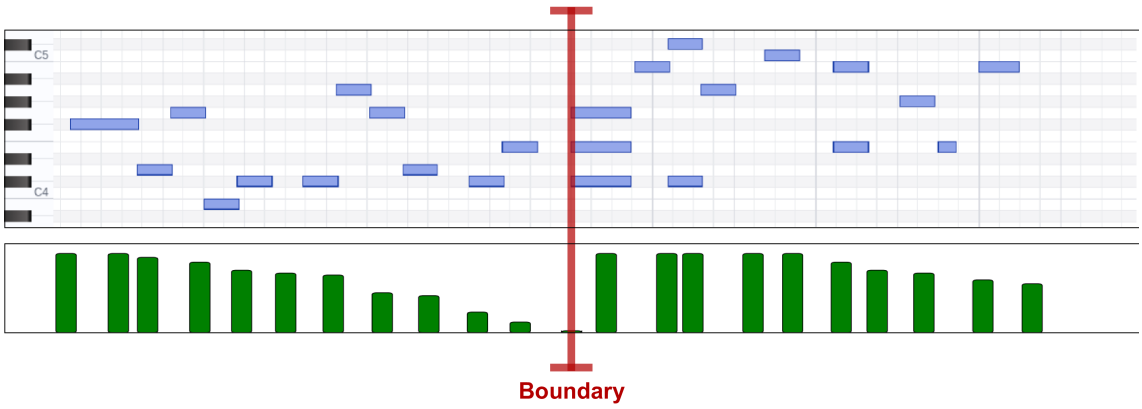


图 4.1：音乐边界及其偏移量的图示。

Wang 等人，2020年），并沿特征维度进行拼接。采用这种架构设计主要基于几个因素。首先，我们在输入层注入边界偏移编码，而不是将其置于 Transformer 主体内部，从而确保核心模型保持不变。这样一来，模型既可以处理带有边界偏移量的输入，也可以处理不带边界偏移量的输入；当缺少边界偏移量时，只需沿特征维度重复学习到的位置编码，即可无缝进行微调。其次，为了区分边界偏移编码和学习到的位置编码，我们避免将二者直接相加。最后，近期研究表明，即使不显式添加位置编码，仅解码器 Transformer 也能通过其内部权重隐式学习位置编码（Haviv 等人，2022年；Kazemnejad 等人，2024）。基于这一发现，我们将学习到的位置编码的特征长度减半，并将剩余特征空间分配给边界偏移编码。由此得到的向量序列由位置编码及其扩展的边界偏移编码构成。按照标准 Transformer 模型的做法，我们将该序列与词元嵌入相加（Vaswani 等人，2017年），然后将其输入具有相对全局注意力的 Transformer 主体（C. A. Huang 等人，2019年）。

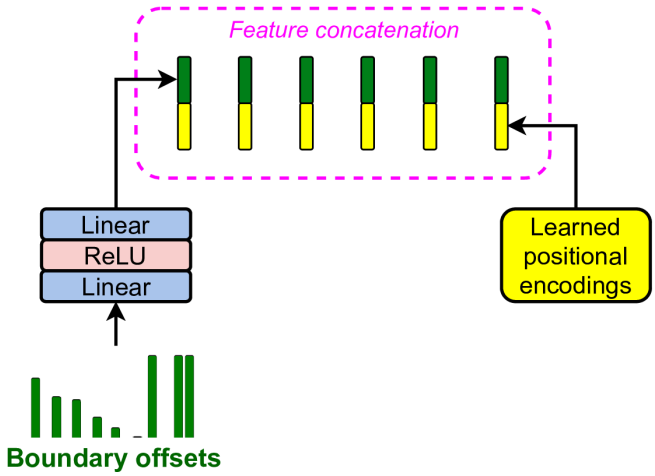


图 4.2：基于边界偏移量的条件机制。

算法 1 概述了推理（即音乐生成）过程中计算边界偏移量的过程。模型以自回归方式生成音乐，每次前向传播生成一个词元。我们通过跟踪生成的TIMESHIFT 词元来维护时间游标，并为每个生成的词元计算边界偏移量，即距离下一个边界所剩的时间。如果模型生成了一个 CHORD 词元，我们就计算时间游标与每个输入边界之间的绝对差值。与时间游标的差值小于预定义灵敏度阈值 1 秒的边界被视为已成功生成。随后，我们将这些边界替换为无穷大，从未来的偏移量计算中移除它们。接下来，无论生成词元的类型如何，我们都会计算其边界偏移量。该偏移量表示到最近的下一个边界的距离。计算方法是从每个输入边界中减去时间游标，并将所有负值（对应于过去的边界）替换为无穷大，以忽略这些边界。然后，我们取所得值中的最小值，作为到最近边界的偏移量，并将其限制在最大值以内，若

算法 1 在推理时于音乐生成过程中创建边界偏移量。

输入：输入边界列表（以秒为单位），$\mathbf{b}$；视频时长，d；效价，x_v；唤醒度x_a**参数**：灵敏度，ξ；到所有边界的距离，$\boldsymbol{\delta}_b$；最大偏移量，δ_{max}；生成的边界掩码，$\mathbf{m_b}$；令牌类型 t_t；令牌值，t_v；时间游标，c；包括前向传播和采样的生成函数，g

输出：生成的词元列表$\mathbf{t}$；边界偏移量列表，$\boldsymbol{\delta}$

令 c $= 0$**;** $\boldsymbol{\delta} = [\,]$**; t** $= [\,]$**. 当 c <d 时执行 # 将词元生成为（类型，值）：** $(\mathbf{t}_t, \mathbf{t}_v) = \mathbf{g}(\mathbf{t}\ \boldsymbol{\delta}, \mathbf{x}_v, \mathbf{x}_a)$ **if** $\mathbf{t}_t$ **== TIMESHIFT then c** $= \mathbf{c} + \mathbf{t}_v$ **else if** $\mathbf{t}_t$ **== CHORD then** $\mathbf{m_b} = |\mathbf{c} - \mathbf{b}| < \xi.\ \mathbf{b}[\mathbf{m_b}] = +\infty$**. 结束 if** $\boldsymbol{\delta}_b = \mathbf{b} - \mathbf{c}\ \boldsymbol{\delta}_b[\boldsymbol{\delta}\mathbf{b} < 0] = +\infty$ $\boldsymbol{\delta}$**.append(min(min(**$\boldsymbol{\delta}$**b)** δ_{max}**))** **t.append((**$\mathbf{t}_t$**,**$\mathbf{t}_v$**)) 结束 while**

所需。生成的边界偏移量和词元会分别追加到相应列表中，以供下一个时间步使用。为简单起见，算法 1 以单个样本为例进行说明，但实际操作中会以小批次执行。训练期间，我们通过构建时间网格而非时间游标，一次性预处理整个输入序列，从而能够同时计算所有词元的边界偏移量。为清晰起见，生成词元的初始列表显示为空；然而，实际操作中我们会以一个开始 词元作为起始。

出现在ON 词元之后、下一个CHORD 词元之前的词元，被视为生成和弦中的音符。为了让生成的和弦更具辨识度，我们提高这些音符的力度。

在训练过程中，真实标注MIDI中和弦的时间位置用作输入边界。正如我们将在第6章中进一步详述的那样，在基于视频的推理过程中，我们会用视频场景切换位置替换这些边界。

接下来，我们将讨论以情绪为条件的音乐生成。

4.2 基于情绪的音乐生成

基于情绪的音乐生成是我们条件音乐生成模型的第二个模块。与时间条件控制相比，基于情绪的条件控制旨在建立视频与音乐之间更高层次的联系，重点关注情感基调。

为了实现基于情绪的音乐生成，我们此前已为 MIDI 数据集获取了情绪标签，如第3章所述。这些标签在模型开发过程中用作条件输入。接下来，我们将介绍我们为基于情绪的音乐生成所开发的方法。

我们模型的基础是音乐 Transformer（C. A. Huang 等人，2019年），它是一种采用相对位置嵌入的仅解码器 Transformer。我们首先设计了一个原始 模型，不使用任何条件化。该模型接收音乐 token，通过嵌入层将它们投影到向量空间，并将输出输入音乐 Transformer。

我们探索了几种基于情感特征对音乐生成过程进行条件化的方法，并将它们称为离散 *token*、连续 *token*和连续拼接。其中，离散 *token* 方法代表了条件序列生成领域的最先进水平（Hung 等人，2021；Keskar 等人，2019年；Payne，2019年），该方法会在原始序列前添加一个离散控制 token——在我们的情形中，即 token 化 MIDI 序列。不过，由于我们的情感条件是连续的，而非离散的，因此我们设计了其他方法来融入这些连续条件，从而能够对生成音乐进行更细粒度的控制。

在 离散 *token* 方法中，我们将连续的情感条件划分到各个分箱中，从而将其转换为离散值。具体而言，我们使用 5 个大小相等的分箱对条件值进行量化，并将中央分箱的索引设为 0。分箱数量的选择旨在对应典型的语言量词，例如 非常低、低、中等、高和非常高。在输入 Transformer 模型之前， valence 和唤醒度的控制标记会被前置到音乐词元中，也就是沿序列维度进行拼接。这种方法的一个关键局限在于，将连续值分箱可能会造成信息损失。

在下一种称为连续 *token*的方法中，我们直接使用原始连续形式的条件值。效价和唤醒度的值分别通过独立的线性层，以生成与音乐 token 嵌入特征长度相匹配的条件向量。随后，这些条件向量会沿序列维度与音乐 token 嵌入拼接，并输入 Transformer。与离散 *token* 方法不同，这种方法保留了情感条件的连续特性，避免了量化带来的信息损失。

我们最终采用的方法是连续拼接，将两个连续条件值组合成一个向量，然后沿序列维度重复该向量，并将其与每个音乐 token 嵌入拼接。这种设计解决了连续 *token* 方法的一个局限：在该方法中，条件向量与单个音乐词元被赋予同等的重要性。我们认为，情感条件会对音乐序列产生全局影响，不同于仅在局部发挥作用的单个音符。通过将条件向量嵌入每个词元，我们确保整个序列始终包含情感上下文，从而在每一步强化其影响。模型的表征如图4.3所示。

对于给定样本，如果不存在条件标签，即为 NaN（非数值），我们就无法将其量化到分箱中，也无法将其投影到向量空间中。在这种情况下，我们采用学习得到的嵌入，而不是投影。由于任何离散标记都会通过嵌入层投影到向量空间中，因此这种方法等价于在模型的词汇表中设置诸如 无_效价和无_唤醒度之类的特殊离散标记。这种方法不仅能够处理缺失数据

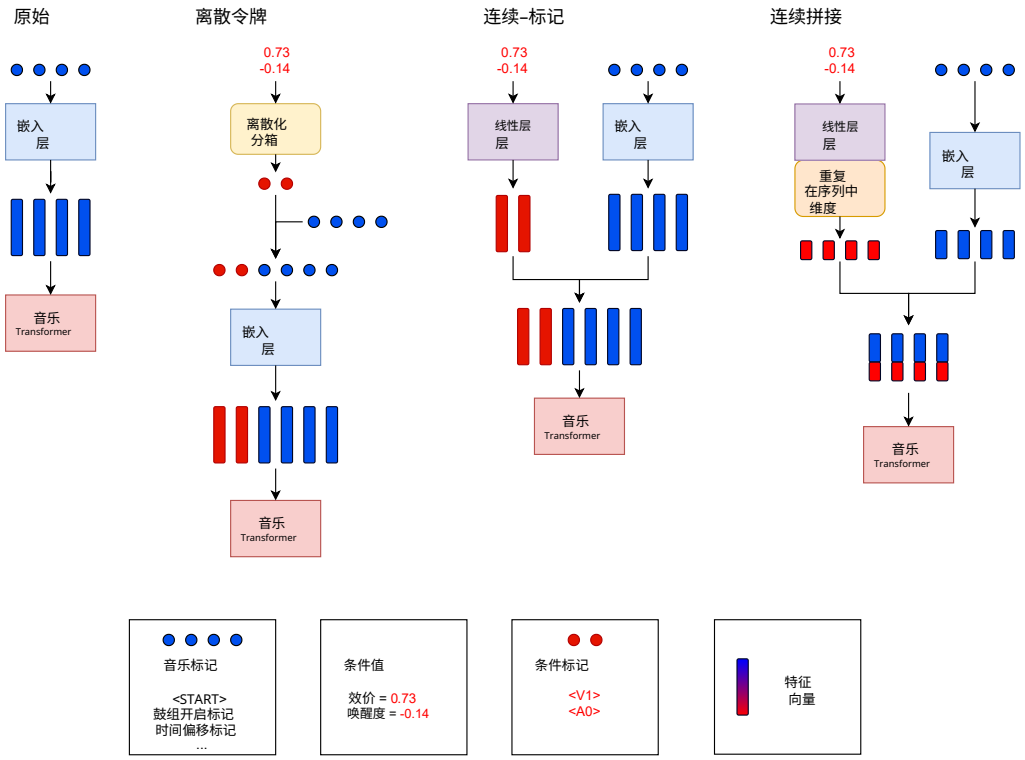


图 4.3：基于情感的音乐生成模型。

还使模型能够在统一框架内切换条件生成和非条件生成。

4.3 数据集与预处理

我们使用 Lakh 钢琴卷帘数据集（LPD）（Dong 等人，2018年）训练音乐生成器。该数据集包含 174,154 个由 Lakh MIDI 数据集（Raffel，2016）衍生的钢琴卷帘。我们使用基于事件的符号音乐表示（Oore 等人，2020年）对钢琴卷帘进行词元化。具体来说，`ON`（音符开始）词元标记音符的起始，`OFF`（音符结束）词元标记音符的结束。这些词元还编码了音高和乐器信息。例如，MIDI 音高为 60（C4）的钢琴音符表示为 `PIANO_ON_60`。我们滤除音高超出钢琴音域（即低于 21（A0）或高于 108（C8））的 MIDI 音符开始与结束事件，因为使用标准 MIDI 音色库无法听到这些音符。由于乐器数量增加会扩大词汇表大小，我们采用 Lakh Pianoroll Dataset-5 变体，其中所有乐器轨道都合并为五个预定义类别：贝斯、鼓、吉他、钢琴和弦乐（Dong 等人，2018年）。不过，我们的方法也适用于乐器分组方式不同的数据集。

我们采用 8 毫秒的时间分辨率，最大偏移量为 1000 毫秒，这与 Oore 等人（2020年）的方法相同。较长的时值用多个连续的时间偏移 词元表示。我们还使用开始 词元标记歌曲的开头，使用小节 词元表示音乐小节，并使用填充 词元统一小批次中的输入序列长度

如前所述，强和弦通过在其第一个开启 词元前插入一个和弦 词元来标记。

由于我们的目标是生成多乐器作品，因此优先选择包含三种或更多乐器的作品。然而，数据集中有许多歌曲仅包含一种或两种乐器。我们没有将其过滤掉，而是为包含两种或更少乐器的歌曲添加前置特殊标记：`FEWER_`乐器；为包含三种或更多乐器的歌曲添加：`MORE INSTRUMENTS`。这些词元_允许用户在推理时指定乐器偏好，同时在训练期间利用整个数据集。

作为情绪标签，我们使用此前构建的 Lakh-Spotify数据集，详见第 3.1 节。我们采用情感维度环形模型（Russell，1980）表示情绪，并略作调整：效价和唤醒度均限制在范围 $[-1, 1]$ 内，因此形成的是方形而非圆形的效价-唤醒度空间。由于 Lakh-Spotify数据集中原始效价值的范围为 $[0, 1]$,，因此需要进行平移和缩放变换。然而，我们在原始效价分布中发现，0.0 处出现了一个意外峰值。检查相应的音乐样本后发现，该值并不总是反映低效价特征，这表明缺失值或未定义值可能被标记为零。因此，我们将这些值替换为 NaN（非数值），最终得到 23,653 个效价标签有效的样本。

为了对唤醒度进行建模，我们采用音符密度和速度等低级 MIDI 特征，遵循 Williams、Kirke、Miranda 等人（2015年）提出的方法；对输出进行定性观察后发现，结果在音乐上连贯，与 H. H. Tan 和 Herremans（2020年）的研究结果相似。我们最初尝试了平均音符密度，取得了一定成效；然而，最终我们仅对含有鼓轨的样本采用估算速度，因为这样得到的定性结果更为一致，详见第 3.1 节。为确保生成过程稳健，我们通过反复试验和人工检查确定了 50 和 150 这两个经验界限，并排除超出界限的速度值。超出这些界限或缺少鼓轨的样本，其唤醒度均赋为非数值。最终，共有 103,735 个样本具有有效的唤醒度值。效价和唤醒度的值均经过平移和缩放，使其落在范围$[-1, 1]$内。最后，为创建测试数据划分，我们按字母顺序排列 LPD匹配子集中的文件名，并将最后 5% 留作测试数据。

4.4 实现细节

训练输入序列是从 MIDI 序列中提取的固定长度数据块。在我们的实验中，我们使用长度为 1216 个词元的输入序列，以最大限度地利用内存。以 0.05 的概率，输入块会从歌曲开头选取，这有助于模型学习歌曲通常如何开始，并更好地理解 `START` 词元。以剩余的 0.95 概率，数据块会从歌曲中的随机位置选取，从而增加输入的多样性。

为了进行数据增强，我们将除鼓以外所有乐器的音高按随机选取的整数值进行移调，该整数值介于 -3 和 3 之间（含两端）。这个相对较窄的范围确保歌曲的情感特征保持不变。通过应用这两种方法，即随机起始点和音高移调，有效训练数据量会显著大于训练数据集划分中的歌曲数量。

我们的模型有 20 层，特征维度为 768。每层有 16 个注意力头，以及一个维度为 3072 的前馈层。在连续拼接模型中，条件向量和词元嵌入的维度分别为 192 和 576，从而确保各模型的 Transformer 输入总维度始终为 768。总体而言，我们的模型约有 1.45 亿个参数。

我们使用 PyTorch 库（Paszke 等人，2019年）实现模型，并在单个 NVIDIA Quadro RTX 6000图形处理器上进行训练。我们使用 Adam优化器（Kingma & Ba，2015年），学习率设为 2e-5。我们的初步实验证实了 Donahue 等人（2019年）的发现：语言建模任务中常见的学习率（约 2e-4）对于 MIDI生成任务来说过高。当训练损失趋于平稳时，我们将学习率降至 2e-6，并继续训练直至收敛。我们采用范数为 1 的梯度裁剪，丢弃率为 0.1，批量大小为 4，输入长度为 1216 个词元。我们使用三角自回归掩码，防止模型关注未来词元。

在推理阶段，生成开始之前，输入序列仅由 `[START,BAR]` 词元组成。或者，为了生成开头更突兀的音乐作品，我们可以省略 `START` 词元，仅使用 `BAR` 词元作为提示序列。这种方法通常会生成一段类似于歌曲中间部分的输出。此外，用户可以选择任意 MIDI 序列作为提示序列，让模型接续给定的旋律。

我们以自回归方式生成输出：将生成的词元追加到输入序列中，形成下一时间步的输入序列。当生成的序列开始超过 1216 个词元的最大长度时，我们仅将生成序列的最后 1216 个词元作为输入，确保不超过长度限制。生成词元时，我们使用核采样，从概率为 $p = 0.7$ 的温度调整后的 Softmax 分布中进行采样（Holtz- man 等，2020年），其中温度设为 1.2。为防止过度重复，如果上一步核采样中的词元数少于 3，我们会略微提高温度。这些超参数通过网格搜索确定：我们测试不同的参数组合，并通过聆听生成的音乐对结果进行人工评估。

现在我们来讨论条件音乐生成的评估方法。

4.5 评估

对我们的时间条件机制进行定量评估具有挑战性。如前文所述，强制模型在条件边界处恰好放置一个和弦，可能会扰乱生成音乐的节奏。相反，我们的目标是引导模型在时间边界附近放置一个和弦，确保其在节奏上与其余生成的

序列相匹配。此外，如果在某个条件边界附近放置和弦会扰乱节奏，我们也希望允许模型忽略该条件边界。因此，这些时间条件属于软约束，而非硬约束。我们的目标并不是最大化准确率指标；相反，在基于视频的音乐生成的上下文中，我们旨在实现主观上的同步感。为此，我们选择进行主观评估，将模型的输出与各种输入视频叠加呈现。相关评估将在第6.3节进一步详述，其中我们通过用户调查收集反馈。调查中的一个问题专门针对时间对齐，询问：“就节奏和时序而言，音乐与视频的匹配程度如何？”该问题有助于评估时间边界条件控制方法的有效性。

我们采用定性和定量方法评估情绪条件控制方法。正如第 6.3 节将详细介绍的那样，我们的基于视频的用户研究包含这样一个问题：“音乐在情感方面与视频匹配得有多好？”在定量评估中，我们通过禁用时间条件控制来单独考察情绪条件机制。我们使用负对数似然（NLL）、Top-1 准确率和 Top-5 准确率这些指标评估情绪条件控制方法。测量 Top-n 准确率时，对于每个词元，如果真实类别位于模型输出概率最高的n 项之中，就认为模型对该词元的预测准确。评估配置与训练相同，即使用长度为 1216 个词元的数据块，并计算目标序列中每个词元的损失。这比给定完整序列后仅预测下一个词元要困难得多，因为在极端情况下，模型只能根据 `START` 和 `BAR` 词元来预测歌曲的第一个音符。我们通过依次选取不重叠分块，确保测试集全部纳入评估，共得到 1836 个分块。

此外，我们还按照 Hung 等人（2021）的方法，通过分析条件模型生成样本的情绪内容，对其进行定量评估。为此，我们首先使用训练数据集划分中的样本训练一个回归模型，以预测样本的情感值。该回归模型采用 8 层音乐 Transformer 架构，最后一层输出两个连续值，即效价和唤醒度。接着，我们使用训练好的条件生成模型，基于一组条件进行推理，然后使用训练好的回归模型预测生成样本的情绪内容。作为误差指标，我们使用回归模型的预测值与推理时输入条件之间的归一化 L_1距离。为了与离散词元模型进行公平比较，我们将条件值设为离散条件词元所用分箱的中点，即 -0.8、-0.4、0、0.4 和 0.8。将 5 个效价值与 5 个唤醒度值组合后，共得到 25 对条件值。对于每个模型和每种条件，我们都生成 8 个样本，不进行“挑选”，并报告平均误差。每个样本包含 4096 个词元。回归模型的输入长度为 1216 个词元，与生成器相近。样本通过重叠率为 50% 的滑动窗口输入回归模型，并对输出取平均值。图 4.4展示了评估生成样本的总体方案。

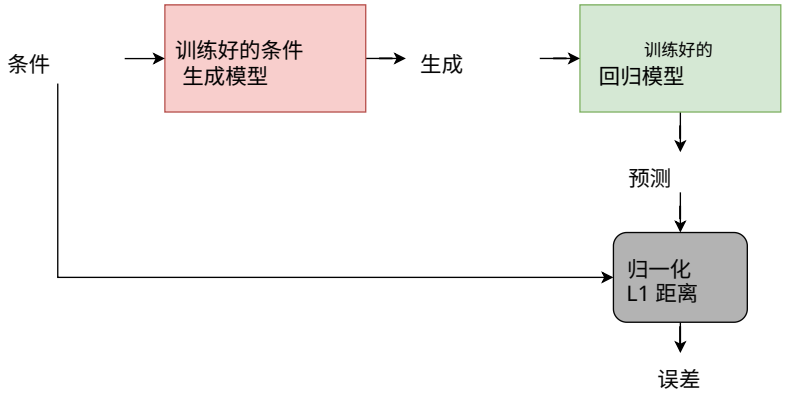


图 4.4：情感表征评估流程。

我们还将生成的样本在线公开[1]。如前所述，我们使用条件模型分别生成 200 个样本。由于生成的旋律具有固定数量的词元，其时值与速度成反比。以最低唤醒度条件生成的旋律速度最慢，平均时长为 159.6 秒。我们还使用原始模型生成了另外 200 个不带条件的样本。总体而言，我们共展示 800 个样本。使用 Fluidsynth 软件和 FluidR3 GM音色库将 MIDI 文件渲染为 MP3格式[2]。_

4.6 结果与讨论

现在，我们介绍情感条件化音乐生成方法的结果与讨论。如前文所述，时间条件机制的结果将在第 6.4节中稍后讨论。

根据基于预测准确率的评估，情绪条件控制方法的演奏表现如表4.1所示。NLL 指负对数似然，数值越低越好。Top-1 和 Top-5 指准确率，数值越高越好。在所有指标上，连续拼接方法都优于其他方法，包括最先进水平的离散 *token*方法。

表 4.1：评估期间我们的方法的演奏表现。

模型	NLL	Top-1	Top-5
原始	0.7445	0.7784	0.9513
离散词元	0.7375	0.7885	0.9536
连续词元	0.7122	0.7895	0.9545
连续拼接	**0.7075**	**0.7913**	**0.9548**

[1]https://www.serkansulun.com/midi[2]
https://archive.org/details/fluidr3-gm-gs

表4.2 展示了基于回归的评估结果。误差指推理过程中输入的条件与训练好的回归模型输出之间的归一化 L_1 距离；该回归模型以生成样本为输入。结果表明，连续拼接方法在传达情感方面优于其他方法。请注意，此处未纳入原始方法，因为它不受任何情感信息的条件控制。

表 4.2：我们的方法在推理期间的演奏表现。

模型	误差
离散 token	0.2164
连续 token	0.1951
连续拼接	**0.1948**

在比较所提出方法的演奏效果时，我们推测，与连续拼接 方法相比，离散 *token* 和连续 *token* 方法的主要缺点在于，它们对条件值和序列令牌一视同仁。虽然序列中的每个词元主要用于进行局部预测（即预测附近的词元），但条件值具有全局影响，因为它们会直接影响整个生成样本。我们提出的连续拼接 方法将条件信息融入 Transformer 输入序列的每个嵌入中，从而最大限度地发挥条件信息的作用。我们提出的两种方法还都支持连续值条件，能够对生成过程进行更精细的控制。

为了确保可复现性并帮助未来的研究，我们将数据集以及用于准备该数据集的代码开源[3]。在自然语言处理社区中，从头开始训练大型 Transformer 是一种罕见的做法，通常会被迁移学习所取代，即对开源预训练模型进行微调。然而，符号音乐生成领域并不存在类似现象。因此，我们还将训练好的模型开源，以便其他研究人员通过迁移学习减少训练所需的时间和资源。据我们所知，我们的方法是文献中规模最大的开源符号音乐生成模型，并且是在规模最大的多乐器符号音乐数据集上训练得到的。

本章使用的条件要么源自 MIDI 数据，要么在推理过程中由人工提供。在第 6章中，我们将展示如何从输入视频中自动提取这些相同的条件，并将其传入我们的条件音乐生成器。不过，在介绍完整的基于视频的系统之前，我们先说明如何从视频中获取这些条件。现在，我们来介绍视频分析方法。

[3] https://github.com/serkansulun/midi-emotion

第5章

视频分析

本章将介绍我们的视频分析方法，这些方法为前文所述的条件音乐生成器生成输入。需要说明的是，我们的总体方案遵循逆向设计：先开发音乐生成组件，再构建视频分析模型，以确保每个阶段的输出都具有音乐上的意义。具体而言，我们从两个互补的角度分析视频：低层级的时间边界和高层级的语义。时间分析用于检测场景切换边界，而语义分析则涉及情感与类型分类。在研究的最后阶段，这些从视频中提取的特征将有助于生成与场景切换相契合、和弦边界鲜明的音乐，同时反映视频的情感基调。

如第 2.5 节所述，场景切换提取是一个已解决的问题。因此，我们将研究重点放在更具挑战性的任务——语义视频分类上。我们的基于视频的音乐生成器依赖于视频情绪分类器；然而，正如第 1.2 节和2.5 节所指出的，视频情感数据集规模有限，这限制了模型的复杂度，并且通常需要使用固定且数量较少的输入帧。为应对这一问题，我们首先设计一个适合这些限制的情感分类器。随后，我们将相同的架构用于预告片类型分类——这一任务拥有多得多的数据，使我们能够探索模型在情感数据集所施加的限制之外的能力。

5.1 场景切换提取

在我们的基于视频的音乐生成器中，我们使用场景切换形式的视频边界，以及高层次情绪。由于场景切换检测是一个成熟的问题，我们使用 FFmpeg 软件完成此任务。当连续帧之间的平均像素差异超过 27% 时，我们就识别出一次场景切换。

为了生成接近场景边界的音乐和弦，我们旨在避免出现过多相近的边界。因此，我们对提取出的场景切换应用差分滤波器，并移除间隔少于 4 秒的场景切换。我们通过反复试验和误差分析以及人工检查确定了这些参数。

我们继续介绍语义视频分类的方法。

5.2 语义视频分类

视频情感分类和预告片类型分类都属于语义视频分类。由于这两项任务都涉及使用视听数据进行预测，因此我们对二者采用了相似的架构。视频情绪分类器被集成到最终的基于视频的音乐生成器中；与预告片类型分类器相比，它采用了最新的更新和调整。对于这些差异出现的地方，我们将予以重点说明。

现有的大多数视频分类方法都使用原始像素和音频作为输入，但正如第 1.2 节和 2.5 节所述，这种高维输入空间会在数据有限时导致过拟合。此外，视频本质上是多模态的，涉及物体、场景、语音、音乐、声音事件和屏幕文字等元素，而所有这些都可能有助于语义理解。然而，使用单一模型捕获所有这些模态是一项重大挑战。

为了解决这一问题，我们使用从各种预训练深度神经网络（DNN）中提取的学习特征。这些特征较为紧凑，有助于降低输入维度并缓解过拟合。由于这些特征来自已经在大型数据集上训练过的模型，因此包含丰富且具有语义价值的视频分析信息。重要的是，我们仅以前向模式使用这些预训练模型，不进行任何微调，从而显著降低内存和计算需求。随后，我们在这些特征之上训练轻量级模型来进行预测，在利用高级表示的同时保持训练效率。下面介绍我们用于特征提取的预训练模型。

5.2.1 特征提取

我们使用以下预训练模型进行特征提取，并提取最终分类层之前一层的激活值；文本除外。这种方法比单独使用最终预测结果提供了更丰富、信息量更大的表征。对于文本，我们首先提取原始输出，然后通过其他文本模型进行处理，同样使用倒数第二层的激活值作为特征表征。

我们注意到，我们使用的某些预训练模型也会分析屏幕上的音乐。尽管我们的主要任务是音乐生成，但我们纳入音乐分析模型，以提升视频分类性能，并保持与最先进方法的竞争力，同时也使我们的分类器能够应用于不涉及音乐生成的任务。此外，我们的视频情绪分类器并非依赖专门用于音乐分析的模型，而是使用一种统一的最先进通用音频分析模型。我们不会尝试将音乐与其他音频来源分离，因为这仍然是一个开放的研究难题（Pimpale 等，2016）。

CLIP 对比式语言-图像预训练（CLIP）是一种用于图像理解的最先进水平模型，通过对互联网（Radford 等人，2021）上的大量图像及其对应说明进行对比学习预训练。预训练模型首先调整图像大小，使较长边达到 224 像素，然后

将图像进行 224×224 中心裁剪。随后，模型将每一帧编码为一个 512 维向量。

音频标签 为提取音频特征，我们使用一个针对音频事件标注预训练的模型（Kong 等人，2020年）。该模型处理时长为 3 秒的音频分块，并输出一个包含 527 个不同标签概率的向量。我们提取分类层之前的激活值，从而为每个音频分块得到 128 维向量。

音乐 由于电影预告片通常包含重要的配乐，因此有必要进一步探究其音乐方面。为此，我们使用一个针对音乐流派分类预训练的模型（Choi 等人，2016）提取音乐特征。该模型处理时长为 22 秒的音频分块，并为每个音频分块输出一个包含 50 个不同标签概率的向量，以及 64 维激活值。

我们仅将 Audiotag 和 Music 模型用于流派分类，之后将它们替换为以下改进且统一的音频分析模型，用于情绪分类。

BEATs 音频变换器双向编码器表征（BEATs）是一种音频分类模型，由声学分词器和分类器组成，二者通过迭代方式进行训练（S. Chen 等人，2023年）。该模型能够对音频事件和音乐流派进行分类。

人脸检测器 使用 Ultralytics 团队的模型检测并提取人脸[1]，，该模型基于 YOLO（You Only Look Once）（Bochkovskiy 等人，2020年）。YOLO 是一种实时目标检测算法，它将图像划分为网格，并使用卷积神经网络预测每个网格单元的边界框和类别概率。此模型随后用于情绪分类，未用于流派分类。

表情分类器 随后将裁剪后的人脸输入面部情绪分类器[2]。该模型是一种视觉 Transformer（ViT）（Dosovitskiy 等人，2021），使用 FER-2013（面部情绪识别）数据集（Barsoum 等人，2016）进行训练。模型接收面部图像，并预测其面部表情属于愤怒、厌恶、恐惧、开心、悲伤、惊讶或中性。此模型随后用于情绪分类，未用于流派分类。

自动语音识别（ASR） OpenAI 的 Whisper 模型处理音频输入，检测语音模式并输出文本，正如在流派分类任务中所做的那样（Radford 等人，2023年）。我们使用该模型的较大版本，它还可以将非英语文本翻译成英语。

[1] https://github.com/ultralytics/ultralytics[2] https://huggingface.co/trpakov/vit-face-expression

光学字符识别（OCR）我们还使用 PaddleOCR 模型对每一帧执行了光学字符识别（OCR）[3]。此外，我们还对生成的输出使用了一个预训练的拼写纠错模型[4]。该阶段的总体输出为文本格式。我们通过引入以下额外的后处理模型来改进 OCR 输出。

语言识别与翻译 由于最终的文本编码模型是在英语文本上训练的，因此我们将非英语光学字符识别文本翻译成英语。我们首先使用语言识别模型[5] 来检测是否需要翻译。该模型基于 XLM-RoBERTa（Conneau 等人，2020年），在多个语言识别数据集上训练而成（Conneau 等人，2018年；Keung 等人，2020年；May，2021）。

随后，我们使用 Facebook 的 NLLB-200（不让任何语言掉队）模型（Costa-jussà 等人，2022年）将非英语光学字符识别文本翻译成英语。该模型结合了 Transformer（Vaswani 等人，2017年）和稀疏门控专家混合模型（Almahairi 等人，2016）层，并使用互联网来源文本数据进行训练，可在 200 种语言之间进行翻译。NLLB-200 模型需要指定源语言，而语言识别模型则会自动检测源语言。作为一种大语言模型，NLLB-200 能够翻译长文本，即使输入中包含拼写错误，也能生成无误差的输出。不过，对于英语光学字符识别输出，我们使用以下校正模型。

拼写纠正 Symspell 软件包提供了拼写检查工具[6]。在这些工具中，分词器可以在缺少空格的句子中分隔单词，这对于对缺少空格的光学字符识别输出进行后处理非常有用。我们还使用拼写检查模型来纠正潜在的拼写错误[7]。该模型是一种 T5 Transformer 语言模型（Raffel 等，2020年），在包含合成拼写错误的数据集上进行训练，因此能够纠正任何英语文本。在获取自动语音识别和光学字符识别的最终文本输出后，我们使用预训练语言模型对其进行编码，以获得特征向量。

DistilBERT 对于流派分类，我们使用预训练语言模型对 OCR 和自动语音识别模型的输出文本进行编码，具体而言，采用 DistilBERT（Sanh 等人，2019年），它是 BERT（来自 Transformer 的双向编码器表示）模型（Devlin 等人，2019年）经过知识蒸馏（Bucila 等人，2006；Hinton 等人，2015年）得到的精简压缩版本。DistilBERT 使用的层数少于 BERT，并通过学习 BERT 的输出来模仿其行为。该模型将输入文本转换为词元（包括单词和子词），并将每个词元编码为一个 768 维向量。

[3] https://paddlepaddle.github.io/PaddleOCR/main/en [4] https://huggingface.co/oliverguhr/spelling-correction-english-base [5] https://huggingface.co/papluca/xlm-roberta-base-language-detection [6] https://github.com/wolfgarbe/SymSpell [7] https://huggingface.co/ai-forever/T5-large-spell

文本情感分类器 对于情绪分类，我们使用另一个经过情感分析训练的语言模型。我们采用 RoBERTa 语言模型（Y. Liu 等人，2019年），该模型在 TweetEval情感分类基准（Camacho-Collados 等人，2022年）[8]上进行过训练。该模型可以将给定文本的情感预测为正面、负类或中性。与前述特征一样，我们使用最终分类层之前的激活值。

介绍完预训练特征后，接下来我们将讨论用于处理这些特征并为情感与类型分类任务生成最终预测的模型。

5.2.2 情绪分类

在此任务中，我们旨在将输入视频归入不同的情感类别。我们将模型命名为 VE-MOCLAP，即使用预训练特征的视频情感分类器。我们首先介绍所用的数据集，以及从中提取预训练特征的方法。

5.2.2.1 数据集与特征提取

我们使用 Ekman-6 数据集，因为它是规模最大的情绪标注数据集之一，其中包含任意类型的用户生成视频（B. Xu 等人，2018年）。我们不使用包含特定类型视频的数据集，例如以人脸或社交互动为重点的视频，因为我们希望最终的基于视频的音乐生成器能够适用于任何类型的视频。Ekman-6 数据集包含 1,637 个短视频，训练集和测试集各占一半。每个视频都标注为六种基本情绪（Ekman，1971年）之一：愤怒、厌恶、恐惧、喜悦、悲伤和惊讶。

我们从每段视频中提取固定数量 n 的帧，以及完整的音频；音频会重采样至 16 kHz 并转换为单声道。面部表情分类器和 CLIP 提取的特征都是向量序列，因为它们的输入是帧序列。如果在单个视频帧中检测到多张人脸，我们会对面积最大的两张人脸的特征取平均值。同样，BEATs 模型会处理由 3 秒音频组成的数据块序列。我们提取 n 个音频数据块，以匹配视频帧的数量。CLIP 和 BEATs 会生成 n 个输出向量，而面部表情分类器生成的向量可能更少，具体取决于每帧中是否有人脸。

自动语音识别模型会根据完整音频生成一个文本块。如果源语言不是英语，它会自动将输出文本翻译成英语。光学字符识别模型会为每个视频帧生成文本块。语言识别模型会处理每个文本块；如果识别出的语言不是英语，文本就会被翻译成英语。如果文本本来就是英语，则会先经过分词器，再经过拼写纠正器。随后，将每帧生成的文本块串接起来，形成一个文本块。自动语音识别和光学字符识别生成的文本会分别输入情感分类器。由于情感分类器对任意长度的文本都只预测一个标签，因此会提取一个向量作为特征。

特征提取后，对于单个视频，我们得到 n CLIP、n BEATs、$k \leq n$ 面部表情、1 个光学字符识别情感特征，以及 1 个自动语音识别情感特征。融合这些预训练特征

[8] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

会带来若干挑战。首先，由于这些特征是使用不同的预训练模型提取的，特征向量的长度（维度）各不相同。其次，当特征向量构成序列时，例如面部表情、CLIP 和 BEATs，其时间长度也可能不同。最后，由于这些特征属于不同的模态，特征向量的内容可能大相径庭。下面介绍我们用于融合预训练特征并生成最终情绪预测的方法。

5.2.2.2 分类方法

为处理多模态特征的取值范围差异，我们首先使用从预训练特征集合中提取的统计量，对每个特征执行最小-最大归一化。为处理不同的维度，所有序列输入特征首先被投影为具有共同维度的查询、键和值（Vaswani 等人，2017年）。接下来，交叉注意力模块（Bahdanau 等人，2015年）利用成对序列特征之间的对应关系。注意力模块还包含随机失活和层归一化，并且能够处理时间长度不同的一对序列。与分类任务中的做法相同，注意力输出沿时间维度取平均，得到单个向量。由于光学字符识别和自动语音识别情感特征本身已是单个向量，经过注意力模块后，每种模态都由单个向量表示。然后，我们将全部五个特征向量沿通道维度拼接，得到表示整个视频的单个向量。最后，将该向量输入线性层，再经过 Softmax 层，为每种情感输出一个概率。

我们的模型如图 5.1所示。圆角和虚线轮廓的模块表示经过训练的模块。带括号的模型按条件使用。其他模块是预训练的特征提取器，在推理模式下使用。Q、K 和 V 分别表示查询、调性和值投影。

5.2.2.3 实现细节和超参数

我们首先以每秒 1 帧的频率提取视频帧，并使用$n = 16$ 个视频帧及相应的音频数据块作为模型输入。鉴于 Ekman-6 数据集规模有限，我们采用较小且固定的输入长度，以降低过拟合风险。在后续数据量更充足的流派分类任务中，我们取消了这一限制。

训练期间，我们从随机位置选取n 个视频帧和音频数据块进行数据增强。测试和推理期间，我们以等间隔提取这些内容，以确保全面的时间表示。我们使用给定的训练集和测试集划分报告分类性能，其中分别包含 819 个和 818 个视频。

我们使用交叉熵损失、批量大小为32、丢弃率为0。5，并使用学习率为1e-5的Adam优化器（Kingma & Ba，2015年）。注意力模块有4个注意力头，维度为512。模型约有1100万个可训练参数。我们将训练集的10%用作验证集，并在验证准确率开始下降时停止训练。

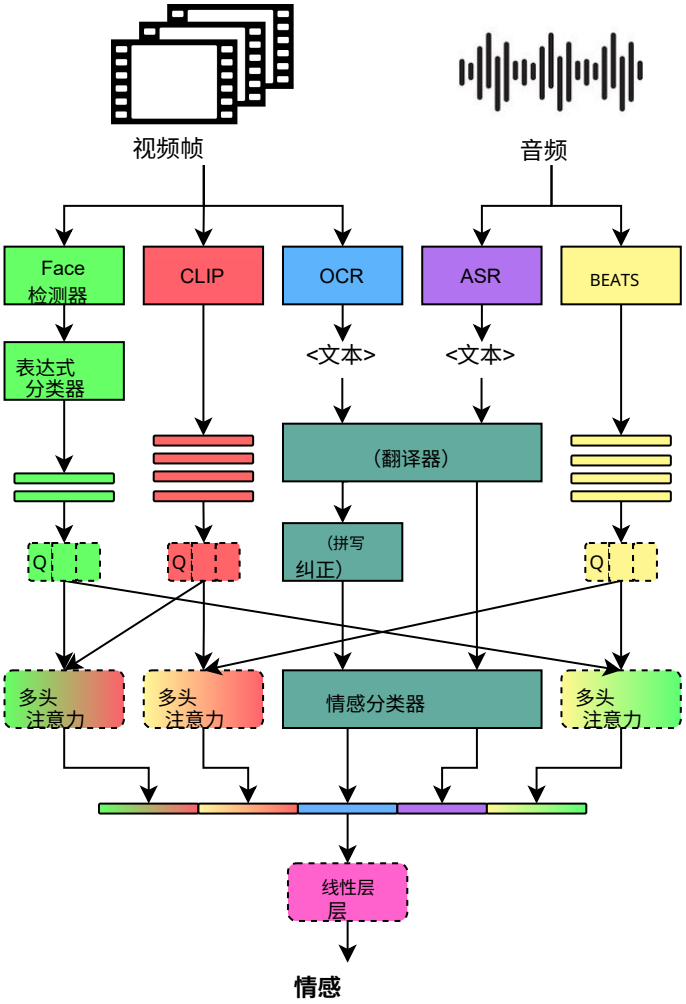


图 5.1：视频情感分类流程。

5.2.2.4 数据集清洗

我们最初检查Ekman-6 数据集时发现了几个有问题的样本。虽然我们报告的是未经编辑的Ekman-6 数据集上的分类结果，但我们也对其进行了清理，以便训练后续第6章中基于视频的音乐生成器所用的模型。Ekman-6 数据集是通过抓取网页上的视频创建的，搜索关键词不仅与情绪类别相匹配，也包含相关词语。我们逐一查看视频以检测有问题的样本。检查文件名后，我们确定了每种情绪类别对应的以下有问题的搜索关键词，并在原文中加了下划线。

愤怒：一个人独自做出令人烦恼的行为，没有其他人在场来感到烦恼或愤怒 y.

厌恶：闪烁的灯光或快速的镜头运动，可能是为了引起眩晕或恶心。还包括与无聊和厌恶感相关的视频。

恐惧：反恐、水下画面、9·11 恐怖袭击 aftermath, 以及嫌疑人抓捕。

喜悦：兜风（驾驶汽车）、音乐《欢乐颂》，以及名为 Joy 的人。

悲伤：沉思

惊讶：分心，以及做出令人印象深刻、被称为惊人的壮举的人。

我们分别从训练集和测试集中识别并移除了 128 个和 130 个有问题的视频。使用清理后的数据，分类准确率提高了 2.6%。然而，我们未将这一结果纳入与最先进水平的比较，因为数据清理改变了测试集的内容，影响了比较的公平性。

5.2.2.5 推理Web应用

作为一项补充工作，我们推出了一款开源网络应用，可对用户提供的视频进行推理，旨在服务广大公众和研究社区[9]。该应用托管在 Google Colab 上，可免费使用图形处理器运行时。用户可以上传自己的视频、提供 YouTube 链接，或使用应用内提供的样本视频。该应用独立完整，开箱即用，无需用户进行任何设置。整个流程简化为 5 个步骤，只需点击鼠标 5 次即可获得结果。连接到图形处理器运行时后，用户应按照以下步骤操作：

步骤 1： 自动下载并提取代码库，然后安装所需的 Python 库。此步骤大约需要 2 分钟。

步骤 2： 下载并构建特征提取器模型和分类器模型。由于这些模型是深度神经网络，此步骤大约需要 3 分钟。请注意，即使要对多个视频进行分类，步骤 1 和步骤 2 也只需完成一次。

步骤 3： 选择输入视频的加载方式。选项包括“示例视频”、“YouTube 链接”和“上传视频”。

步骤 4： 根据步骤 3 中的选择，用户随后选择具体的视频。步骤 3 和步骤 4 只需几秒钟即可完成。

步骤 5： 从输入视频中提取帧和音频，运行预训练的特征提取器，最后运行情感分类器。输出包括带有情感分析结果的自动语音识别（ASR）文本、带有情感分析结果的光学字符识别（OCR）文本，以及 BEATs 音频分类器的预测结果。此外，还会显示一个示例帧，其中展示检测到的人脸及预测的情绪表情、检测到的 OCR 文本框，以及由 CLIPCap 生成的说明文字（Mokady 等人，2021）。请注意，该示例帧仅用于演示，最终情绪分类会使用全部 n 帧。对于 60 秒的视频，此步骤大约需要 30 秒。

5.2.2.6 结果与讨论

表 5.1展示了使用所提供的训练集和测试集划分，在 Ekman-6 数据集上对我们模型的定量性能评估结果，表明我们的方法比最先进水平高出 4.3%。

[9] https://www.serkansulun.com/app

Ta表 5.1：与最先进水平相比，Ekman-6 数据集上的分类准确率。

方法	准确率（%）
ITE（B. Xu 等人，2018年）	51.20
CFN（C. Chen 等人，2016）	51.80
MART（Z. Zhang 等人，2024）	53.17
VAANet（S. Zhao 等人，2020年）	55.30
CTEN（Z. Zhang 等人，2023年）	58.20
KeyFrame（Wei 等人，2021）	59.51
LRCANet（Yi 等人，2024）	59.78
FAEIL（H. Zhang 和 Xu，2023年）	60.44
TAM（Pan 等人，2022年）	61.00
VEMOCLAP（我们的方法）	**65.28**

图 5.2展示了我们在测试集上的分类结果混淆矩阵。在六种情绪中，厌恶的分类准确率最低。这可能是因为，与其他情绪不同，厌恶很大程度上受到社会制约的影响，并受文化道德价值观塑造（Haidt 等，1997年）。被标记为悲伤的视频中，有 20% 被误分类为恐惧，这可能是因为这两类情绪都具有相似的负面情感基调。同样，被标记为喜悦的视频中，有 16% 被误分类为惊讶。根据我们的人工检查，这可能是因为惊讶类别中的大多数视频都传达出积极情感——其中的人物往往不仅感到惊讶，也感到喜悦。

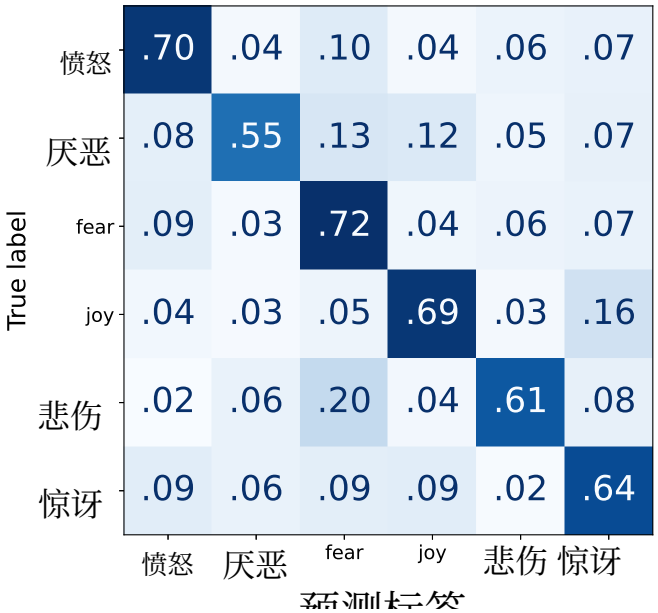


图 5.2：Ekman-6 数据集测试集上按真实标签归一化的混淆矩阵。

据我们所知，我们的推理Web应用是首个用于视频情感分类的点击即运行软件。我们还将训练好的 VEMOCLAP 集成到了视频音乐生成流程中，具体内容将在第 6章中说明。

5.2.3 流派分类

接下来，我们将继续探讨另一项语义任务，即流派分类。在这项任务中，我们利用一个大规模电影数据集，挑战输入帧数量方面的限制。

5.2.3.1 数据集和特征提取

我们使用 MovieNet 电影数据集，其中包含 37.5 万部电影的元数据，以及其中 3.3 万部电影的 YouTube 预告片链接，因此这是目前规模最大、带有标签的预告片数据集。重要的是，该数据集采用多标签标注，这意味着单个视频可以对应多个类型。

我们提取 CLIP、光学字符识别、自动语音识别、音频标签和音乐特征，如图所示 5.3。与情感分类流程不同，我们使用通用语言模型 DistilBERT 处理文本特征，该模型会生成一个特征向量序列，每个词元（词或子词）对应一个向量。

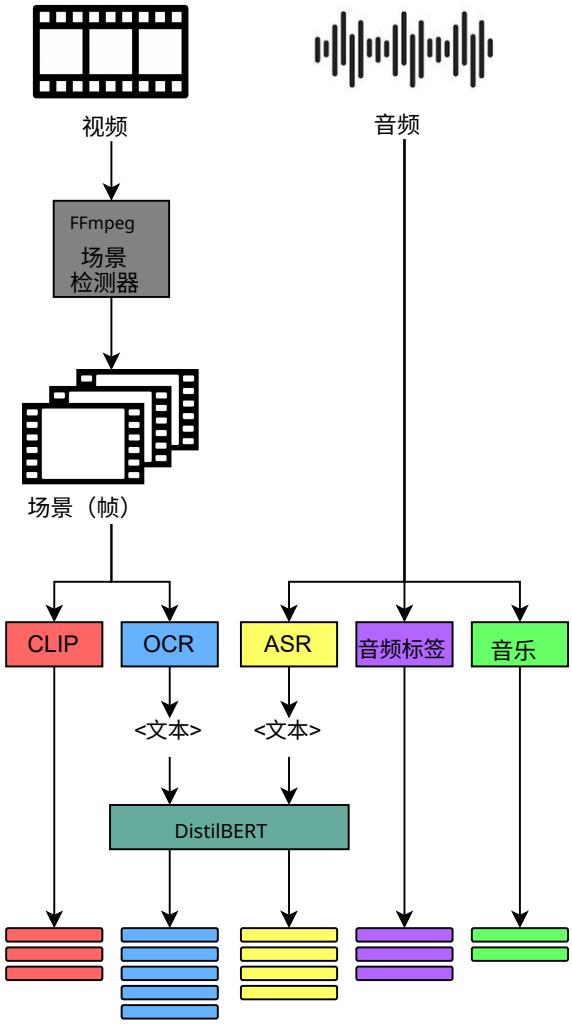


图 5.3：用于流派分类的视频特征提取流程。

5.2.3.2 分类方法

我们构建了多个模型，将先前提取的给定视频特征作为输入，并预测该视频的类型。组合不同特征颇具挑战，尤其是编码向量在通道和时间维度上的长度各不相同。例如，音频事件和音乐特征的向量长度分别为 128 和 64。此外，由于音频事件和音乐网络分别处理长度为 3 秒和 22 秒的音频片段，因此对于同一视频，前者对应的嵌入向量更多。最后，同一特征的向量数量在不同视频之间也有所差异。我们通过采用不同的解决方案和不同的分类模型来应对这些问题。

我们构建并训练三种不同的模型，以融合并分类预训练特征：多层感知机（*MLP*）；整合所有模态特征的单*Transformer* 模型；以及多变换器 模型，其中各个Transformer分别处理来自特定模态的特征。作为基线，我们还实现了一个更简单的模型，该模型不使用多样化的预训练特征，而是使用两个分支分别独立处理原始视频帧和音频。

我们所有模型的最后一层都是大小为21的全连接（FC）层，输出属于21种不同类型的概率。该层之后接一个Sigmoid层，以确保每个输出都是介于0和1之间的概率。

多层感知机 我们首先实现一个简单的多层感知机分类器，参照 Q. Huang 等人（2020年）采用的最先进方法。该方法处理来自若干短片段的原始视频和音频，而我们使用从整个视频中提取的预训练特征。由于多层感知机需要定长输入，我们采用类似的策略，按时间对各模态的特征向量取平均。随后将这些平均向量拼接起来，并输入多层感知机。我们的架构如图 5.4所示。

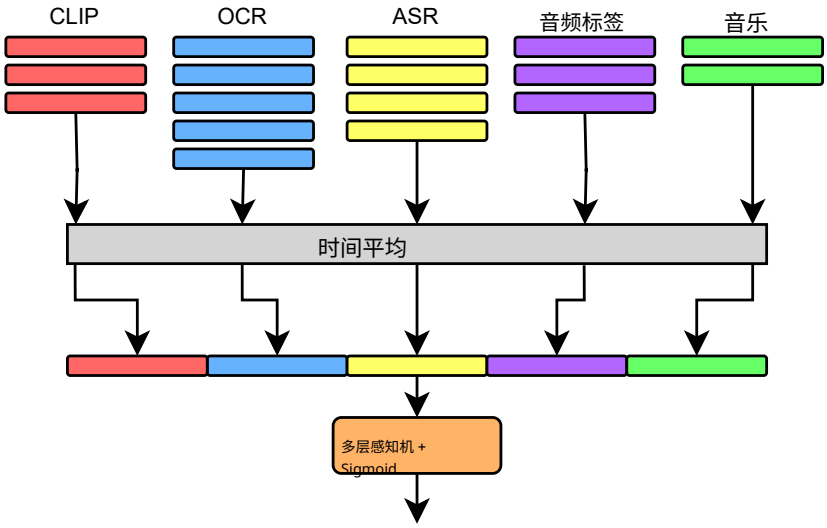


图5.4：我们的多层感知机模型。

在另外两个模型——单Transformer和多变换器——中，我们会选择对 OCR 和自动语音识别等文本特征进行时间平均。CLIP、音频标签和音乐等非文本特征取自最终预测层之前的激活值；而 OCR 和自动语音识别特征则是将 DistilBERT 模型应用于识别出的文本而得到的。识别文本中的任何错误都会传递到 DistilBERT 模型中，可能导致输出特征失真。为减少由此产生的噪声，我们尝试沿时间维度对 OCR 和自动语音识别等文本特征取平均。

单Transformer 为减轻沿时间维度对特征进行平均所造成的显著信息损失，我们采用 Transformer 模型来捕捉视频序列中的短期和长期对应关系。我们将序列分类器用于 Transformer：在序列前添加一个特殊的可学习向量，称为 `<CLS>`向量。Transformer 会为输入序列中的每个元素生成输出向量，但在分类任务中，只有与 `<CLS>` 向量对应的输出向量会传递到下一层，其余向量则被丢弃（Devlin 等人，2019年）。

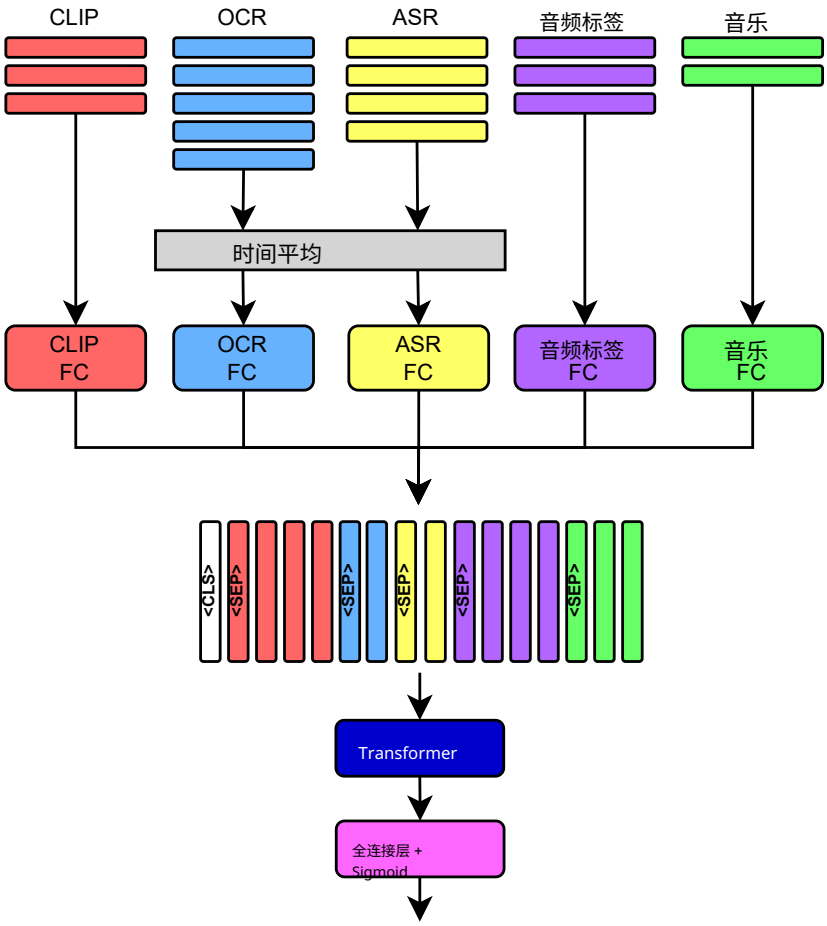


图 5.5：我们的单变换器模型。

虽然Transformer能够处理长度不一的序列，但序列中的向量在通道维度上的大小仍须相同。为此，我们使用全连接层将不同模态的特征向量转换为相同大小。需要指出的是，该层会分别作用于序列中的每个向量，因此不会改变其输入序列中的向量数量。接下来，我们沿时间维度拼接所有向量，再将它们输入单个Transformer。这种方法旨在将所有特征作为一个序列进行处理，从而利用不同特征在时间层面的对应关系。不过，使用单个Transformer 块和一组权重处理不同模态，也会带来处理复杂度方面的权衡。

为确保 Transformer 能够区分来自不同模态的特征向量，我们对每种模态的序列分别使用独立学习的位置嵌入（B. Wang 等人，2020年）。此外，我们还使用分隔符（<SEP>）向量，即对 <SEP>词元进行编码后得到的向量（Devlin 等人，2019年）。这些是针对每种模态单独学习的向量，并添加到每个序列的开头。最后，与几乎所有分类器 Transformer 一样，我们在输入序列前添加一个可学习的分类（<CLS>）向量，即对<CLS> 词元进行编码后得到的向量（Devlin 等人，2019年）。与该分类向量对应的输出向量，即第一个向量，被视为 Transformer 的预测结果，并输入最终的全连接层，而 Transformer 输出序列中的其余向量则被舍弃。总体模型如图5.5所示。

多变换器 鉴于使用单个 Transformer 处理多模态特征可能会因输入复杂度增加而效率低下这一假设，我们设计了一个最终模型，为每种模态分别采用独立的 Transformer 模型，如图 5.6所示。所有 Transformer 的输入前均添加 <CLS> 词元向量，相应的输出向量沿通道维度拼接后，再输入全连接层以获得最终概率。该方法能够利用不同特征在全局层面的潜在对应关系，但无法利用它们在时间层面的对应关系。

视觉 Transformer 基线模型 为了评估我们使用多模态预训练特征的方法，我们还实现了一个强大的基线模型进行比较。与主要模型不同，这个基线模型并未整合多种预训练特征，但仍利用了强大的预训练神经网络以及额外的可训练层。它处理由二维原始视觉帧和音频频谱图组成的序列，而先前介绍的模型处理的则是由一维预训练特征组成的序列。

我们的基线模型使用视觉 Transformer (ViT) 将每个视频帧划分为图像块，对其进行编码，并使用标准 Transformer 架构处理由此得到的向量序列（Dosovitskiy 等人，2021）。对于音频，该模型采用音频频谱图 Transformer (AST)，这是一种 ViT 变体，专为直接处理音频频谱图而设计（Gong 等人，2021）。尽管 ViT 完全基于注意力机制，不包含任何卷积层，但其表现优于卷积神经网络，在图像（Dosovitskiy 等人，2021）、视频（Arnab

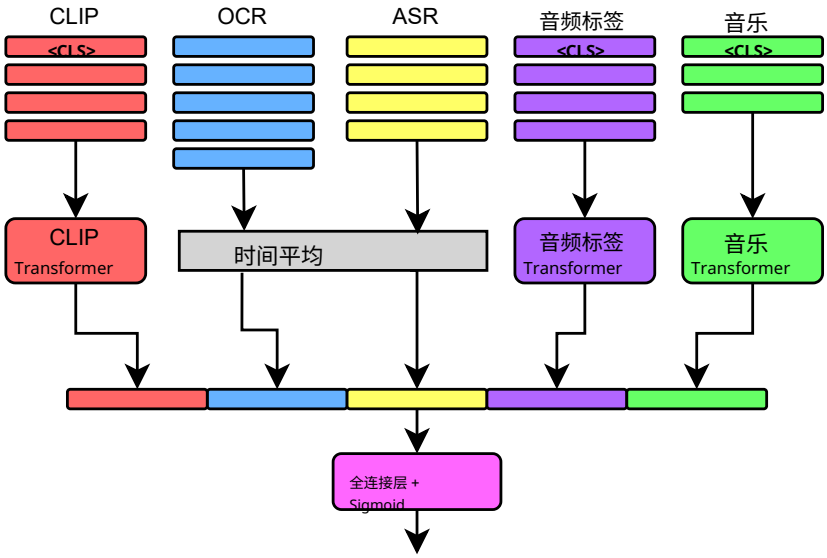


图 5.6：我们的多变换器模型。

等人，2021）和音频（Gong 等人，2021）分类任务中取得了最先进的结果。

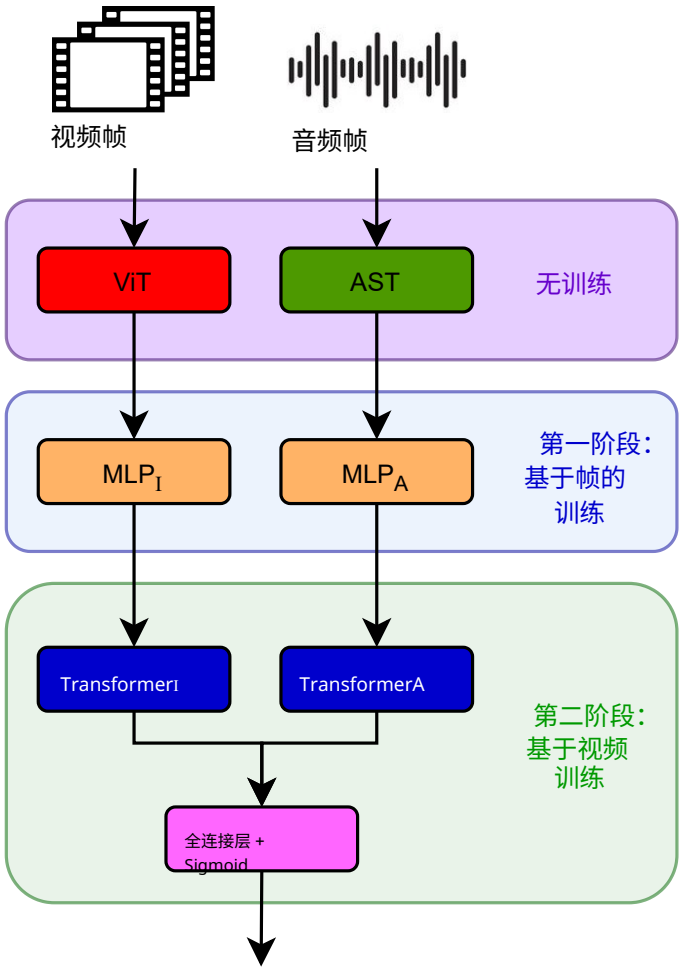
虽然我们的主要目标是视频分类，但将视频视觉 Transformer（ViViT）（Arnab 等人，2021）直接应用于我们的任务并不合适，主要有两个原因。首先，ViViT 不包含音频，而音频是我们任务的关键部分。其次，我们的数据集包含不连续的帧，这些帧呈现不同的场景，彼此之间没有视觉连续性。因此，与其像 ViViT 那样拼接来自不同帧的图像块，不如单独处理这些帧。

ViT 和 AST 模型分别在 ImageNet（Deng 等人，2009）和 AudioSet（Gemmeke 等人，2017年）数据集上进行了预训练；它们分别是用于图像分类和音频分类的大规模基准数据集。在初步实验中，我们尝试训练整个模型，包括对 ViT 和 AST 进行微调，但发现出现了严重的过拟合，且无法通过随机失活（Srivastava 等人，2014年）等标准正则化技术加以缓解。为避免这一问题，我们冻结 ViT 和 AST 的参数，并将它们的输出层替换为可训练的多层感知机，分别命名为图像多层感知机和音频多层感知机。

预训练的 AST 专为处理按 10 秒片段组织的音频频谱图而设计，我们将这些片段称为“频谱图帧”。为区分两者，我们用“图像帧”描述上下文中的视觉场景。由于数据集中的预告片时长超过 10 秒，我们将完整的音频频谱图切分为 10 秒的数据块，重叠率为 50%。类似地，ViT 处理单独的图像帧，而我们的预告片由多个帧组成。因此会得到一个向量输出序列，其中每个向量对应单独的图像帧或音频帧。为了处理这些序列并获得最终预测，我们构建了一个融合模块。为此，我们对图像序列和音频序列分别使用标准 Transformer 分类器，为每种模态生成一个向量。最后，我们将这两个向量拼接起来

a并使用线性层和 Sigmoid 层进行处理，以获得最终的标签预测 .

即使冻结预训练ViT和AST模型的参数，以端到端方式训练整个模型仍会导致过拟合。这是因为我们的训练数据集规模相对较小，仅包含 33k 个样本；而 ViT 和 AST 的原始作者则使用规模庞大的数据集训练模型，例如包含 21M 个样本的 Imagenet（Deng 等，2009年）和包含 2M 个样本的 AudioSet（Gemmeke 等，2017年）。为解决这一过拟合问题，我们采用两阶段训练方法。首先，分别使用单独的帧训练图像和音频多层感知机。在此阶段，目标是这些帧所属视频的标签。接下来，冻结多层感知机的参数，并使用完整视频训练融合模块。基线模型及其训练方案如图 5.7所示。



Figu图 5.7：我们的基线模型。下标“I”和“A”分别指图像和音频 y.

5.2.3.3 实现细节和超参数

我们通过筛选过程，从数据集中剔除时长过长和过短的视频。具体而言，我们计算视频时长的四分位数，并剔除超出内围

范围的样本，即时长短于 $Q1-1.5\times IQR$(对应 19.6 秒）或长于$Q3+1.5\times IQR$(对应 214.4 秒）的样本。经过筛选，原有的 32647 个视频中保留了 26412 个。此外，按照 MovieNet 论文（Q. Huang 等人，2020年）中的方法，我们将标签限制为出现频率最高的 21 种类型，即动作、冒险、动画、传记、喜剧、犯罪、纪录片、剧情、家庭、奇幻、历史、恐怖、音乐、歌舞片、悬疑片、爱情片、科幻片、体育片、惊悚片、战争片和西部片。

为确保数据集划分公正，我们按照 YouTube ID 的字母顺序排列视频，因为这些 ID 是随机生成的。遵循原始 MovieNet 论文中的方法，我们将数据集划分为训练集、验证集和测试集，比例分别为 0。7、0。1 和 0。2，对应的样本数分别为 18488、2641 和 5283（Q. Huang 等人，2020年）。我们通过网格搜索确定超参数，以验证集上的平均精度均值最高为优化目标。测试集仅用于报告最终结果。

平衡精确率和召回率的一个直接方法是调整决策阈值，但这样无法与文献中的研究进行比较，因为大多数研究在报告结果时都将决策阈值固定为 0。5（Q. Huang 等人，2020年）。另外，在训练过程中，我们可以通过为正向标签对应的损失赋予恒定权重来平衡精确率和召回率。对于单个样本，加权二元交叉熵损失为：

$$Loss=-\frac{1}{C}\sum_{c=1}^{C}w\cdot y_c\cdot\log(p_c)+(1-y_c)\cdot\log(1-p_c)$$

其中，w 表示正向标签的权重，C 表示类别总数。为了使精确率和召回率值与 MovieNet 中观察到的值保持一致，我们采用 0.25 的权重。

Transformer 架构和多层感知机架构的平均模块都可以处理长度不一的序列。然而，在训练过程中，我们使用定长输入，以便进行小批量处理。较长的序列会被截断至所需长度，而较短的序列则使用零向量进行填充。每种预训练模态的特征向量长度通过探索性数据分析确定，具体而言，是使用箱线图分析所有样本中的序列长度。最大序列长度设定为上相邻值（箱线图的上须），即第三四分位数以上四分位距的 1.5 倍。因此，各模态的序列长度分别设定为：CLIP 为 216，光学字符识别为 64，自动语音识别为 86，音频标签为 140，Musicnet 为 18。使用单变换器模型时，拼接后的序列总长度为 524。需要注意的是，在对单个样本进行推理时，我们的模型可以处理长度不一的输入，无需进行填充或截断。

对于多层感知机模型，共有1层，模型维度为256，参数总量为57k。单变换器模型由2层、8个注意力头组成，模型维度为256，参数总量为8.56M。多变换器模型采用各自独立的Transformer，每个具有1层、8个注意力头，模型维度为128，总计6.98M

参数。就基线模型而言，图像多层感知机和音频多层感知机均有2层，维度为768。各融合变换器均有1层、4个注意力头，维度为768。所有Transformer均采用范数为1的梯度裁剪进行训练。模型使用Adam优化器（Kingma & Ba，2015年）进行训练，学习率为1e-5，丢弃率为0.5，批量大小为32。这些模型使用PyTorch库（Paszke等人，2019年）实现，并在一块具有24 GB显存的NVIDIA Quadro RTX 6000图形处理器上训练。

5.2.3.4 结果与讨论

我们首先将本模型的性能与基线模型以及 MovieNet 论文中提出的模型进行比较（Q. Huang 等人，2020）。推理时，每次对测试集中的一个样本（视频）进行处理，使用每段视频的完整时值，不对序列进行填充或截断。按照相关文献的惯例，我们报告按各个标签（类型）计算并取平均值的宏平均精确率、召回率和平均精度均值（mAP）。

表 5.2展示了本模型与相关文献中模型的总体性能比较。在该表中，P、R 和 mAP 分别表示精确率、召回率和平均精度均值。精确率和召回率的决策阈值设为 0.5。最先进模型的结果取自 MovieNet 原论文（Q. Huang 等人，2020）。为便于理解，缩写模型的全称如下：TSN（时间片段网络）（L. Wang等人，2016）、I3D（双流膨胀三维卷积网络）（Carreira & Zisserman，2017年）以及 TRN（时间关系网络）（B. Zhou、Andonian等人，2018年）。

我们的基线模型在召回率和平均精确率方面优于文献中的模型。使用预训练特征的模型在所有指标上均优于其他模型。我们表现最佳的模型——多变换器——显著提升了当前最先进水平。除精确率外，它在各项指标上均优于其他模型；在精确率方面，其表现略逊于我们的多层感知机模型。

表 5.2：我们的模型与当前最先进水平的性能比较。

模型	P@0.5	R@0.5	mAP
TSN	78.31	17.95	43.70
I3D	69.58	16.54	35.79
TRN	77.63	21.74	45.23
MovieNet	79.74	24.97	46.88
基线	73.78	32.17	59.73
MLP	**82.05**	33.51	63.16
单Transformer	81.00	37.11	65.09
多变换器	82.00	**38.33**	**66.02**

在表 5.3 中，我们详细分析了所有类型的结果。为简洁起见，我们仅展示文献中最佳模型 MovieNet (MN)（Q. Huang 等人，2020年）、我们的基线 (BL) 模型，以及我们表现最佳的多变换器 (MT) 模型。P、R 和平均精度均值

分别表示精确率、召回率和平均精度均值。对于精确率和召回率，0.5 被用作决策阈值。

表5.3：各类型结果及宏平均值。MN：MovieNet（Q. Huang等人，2020年），BL：基线（我们的方法），MT：多变换器（我们的方法）。

	精确率@0.5			召回率@0.5			mAP		
	MN	BL	MT	MN	BL	MT	MN	BL	MT
动作	73.96	86.59	**87.37**	22.21	40.48	**48.12**	54.60	75.64	**79.20**
冒险	75.24	74.71	**77.51**	24.72	22.48	**28.67**	53.06	56.76	**60.93**
动画	93.16	**96.38**	96.19	74.09	86.59	**92.28**	86.45	94.65	**96.18**
传记	**100.00**	53.85	70.00	0.04	2.86	**5.71**	9.13	24.29	**36.68**
喜剧	68.61	88.55	**90.39**	48.65	**57.94**	57.05	68.81	85.47	**87.84**
犯罪	74.12	72.00	**80.22**	**39.30**	8.33	22.53	49.25	50.81	**60.25**
纪录片	85.49	91.30	**92.37**	4.79	70.87	**81.71**	21.03	90.29	**94.64**
剧情	71.16	85.70	**89.40**	**79.42**	49.16	55.06	79.95	83.08	**86.77**
家庭	82.55	83.08	**86.58**	27.11	40.14	**48.08**	52.19	67.09	**73.60**
奇幻	69.83	79.37	**87.88**	13.51	12.02	**20.91**	39.12	48.78	**59.59**
历史	82.90	63.16	**90.00**	**12.52**	6.49	9.73	34.41	33.63	**41.05**
恐怖片	70.03	86.49	**88.38**	8.76	51.04	**55.82**	35.51	79.48	**84.24**
音乐	**89.04**	66.67	68.06	27.24	20.48	**29.52**	47.13	42.17	**48.36**
歌舞片	73.58	70.59	**75.00**	4.45	11.21	**14.02**	22.88	32.86	**41.27**
悬疑片	**76.42**	0.00	57.14	**7.76**	0.00	1.20	**39.70**	25.28	30.68
爱情片	71.93	76.47	**82.32**	14.02	13.65	**15.75**	49.27	51.16	**58.59**
科幻片	81.35	72.09	**83.60**	14.51	29.67	**37.80**	44.14	54.11	**68.00**
体育片	**94.97**	69.51	75.86	21.99	42.86	**49.62**	39.59	60.44	**66.80**
惊悚片	64.98	**79.35**	78.17	14.50	28.79	**37.82**	49.80	69.54	**71.76**
War	**86.27**	64.71	75.86	12.80	20.89	**27.85**	34.41	47.66	**56.40**
西部片	88.89	88.89	**89.80**	51.93	59.70	**65.67**	73.99	81.14	**83.53**
平均值	79.74	73.78	**82.00**	24.97	32.17	**38.33**	46.88	59.73	**66.02**

在全部21种类型中，我们的模型在精确率方面优于其他模型的有14种，在召回率方面有16种，在平均精度均值方面有20种。传记、历史和悬疑片等类型的召回率极低等异常值，可归因于MovieNet数据集中固有的不平衡（Q. Huang 等人，2020年）。我们有意在训练前不对数据进行平衡，以便与MovieNet论文中提出的分类模型进行公平比较。我们还认为，该数据集密切反映了电影类型在现实世界中的分布，准确体现了传记、历史和悬疑片等类型相对稀少的情况。

在表5.4 中，我们使用表现最佳的多变换器模型进行了消融研究，展示了添加各项特征所带来的平均精度均值提升以及运行时间增加，并与基线模型进行了比较。第一行显示的是基线模型的演奏。星号（$*$）表示对序列（时间）维度上的特征取均值。运行时间定义为推理期间处理单个视频的平均时长（秒），包括提取其预训练特征和进行分类。

Tab表 5.4：加入预训练特征对平均精度均值和运行时间的影响 .

CLIP	音乐	音频标签	OCR	ASR	mAP	运行时间（秒）
					59.73	7.08
✓					64.73	5.32
✓	✓				65.17	5.76
✓	✓	✓			65.31	5.95
✓	✓	✓	✓		63.33	25.85
✓	✓	✓	*		65.46	31.52
✓	✓	✓	✓	✓	64.66	31.44
✓	✓	✓	*	*	**66.02**	33.59

虽然加入 OCR 和 ASR 等文本特征起初似乎会降低性能，可能是由于文本预测错误，但沿时间维度对这些特征取均值可以减少噪声，并提升平均精度均值。加入文本特征也会显著增加运行时间。这主要是因为 CLIP、Music 和音频标签等模型旨在预测单个嵌入向量或标签，而 OCR 和 ASR 模型则预测文本序列；这可以视为以自回归方式预测对应于每个词和子词的多个标签。但最重要的是，不包含文本特征的设置相比基线模型能够取得更好的分类性能和更快的运行速度。此外，我们的分类模型可以无缝整合由未来开发的新型且可能更高效的 ASR 和 OCR 模型生成的特征，从而进一步缩短总体运行时间。

最后，在图 5.8 中，我们尝试进一步解释模型取得成功的原因。文献中的模型会从每个视频中随机抽取少量固定数量的帧用于分类。MovieNet 基线（Q. Huang 等人，2020年）仅使用 8 个场景。此外，它们使用无法处理序列的多层感知机，因此需要沿时间维度对这些少量帧的特征进一步求均值，从而造成额外的信息损失。我们通过一项简化实验，展示输入模型的帧数的重要性。我们仅使用 CLIP 特征，分别采用从每个视频随机位置提取的 8、16、32、64、128 和 256 帧。我们只使用 CLIP 特征，因为不同预训练特征的特征向量数量有所不同，而且难以找到与特定数量的 CLIP 特征相对应的确切特征数量。

由于我们使用单一输入模态，我们的单Transformer和多变换器架构都采用标准的Transformer分类器。因此，我们使用多层感知机和Transformer处理CLIP特征序列。对于多层感知机，我们按照（Q. Huang 等人，2020年）的方法，对不同帧的特征沿时间维度取平均。相反，Transformer 模型能够处理序列，因此我们将每一帧的特征原样输入。

从图5.8中可以看出，使用更多帧始终能提升演奏效果。此外，随着帧数增加，Transformer 相较于多层感知机的优势逐渐显现，这得益于其能够无缝处理长序列。

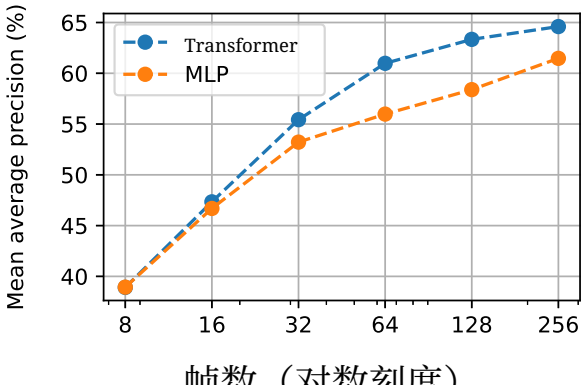


图 5.8：仅使用 CLIP 特征时，帧数与平均精度均值的关系。

总之，我们的流派分类模型之所以成功，是因为结合了多种预训练特征，利用了所有视频场景和完整音轨，并避免了时间平均。相反，我们将完整的特征序列输入 Transformer，而它擅长处理长序列。这些设计选择得以实现，离不开大规模的 MovieNet 数据集。

我们关于电影流派分类的研究凸显了提升视频情感分类效果的潜力，尤其是未来出现带有情绪标签的大规模视频数据集之后。目前，由于现有视频情感数据集规模有限，我们的基于视频的音乐生成系统中的情感分类器使用一组固定且规模较小的输入帧。

现在，我们已准备好整合所有组件，构建基于视频的音乐生成器。该系统将视频情感分析和时间分析（通过场景切换检测实现）模块，与同时采用情感和时间条件控制的音乐生成器结合起来。最终的 MIDI 输出通过 Fluidsynth 软件合成为音频。（提醒一下，引言中介绍的图1.1展示了完整的系统架构。）我们将在下一章详细介绍整合过程。

第6章

基于视频的音乐生成

本章将说明如何整合所有组件，以构建基于视频的音乐生成器。具体而言，我们将整合两阶段双分支流水线中的四个模块：用于音乐生成的时间边界条件器（第 4.1 节）、用于音乐生成的情感条件器（第 4.2 节）、视频场景切换（边界）提取器（第 5.1 节），以及视频情感分类器（第 5.2.2 节）。我们将模型命名为 *EMSYNC*，因为它通过匹配音乐和视频的**情**感，并同步两者的时间边界，实现音乐与视频的对齐。

在继续之前，我们先解决视频情绪分类器（VEMOCLAP）与基于情感的音乐条件控制之间情绪表征不兼容的问题。VEMOCLAP 输出离散情感类别的概率分布，而音乐生成器则以连续效价和唤醒度值所表征的情绪作为条件，取值范围为 -1 至 1。

6.1 情感映射

我们在视频分析与音乐生成模型的情绪表征之间建立对应关系。具体而言，我们将离散情绪映射为连续效价-唤醒度值。

Russell 和 Mehrabian（1977）开展了用户研究，以确定离散情感类别对应的效价和唤醒度值。他们将研究结果列于一张表中，表中包含每种类别型情绪的效价和唤醒度均值与标准差。我们提取并使用与 Ekman 的六种基本情绪对应的值，如表6.1所示（Ekman，1971年）。

利用视频分类器的输出概率，以及各情绪对应的均值和标准差，我们构建高斯分布混合。用户可以从混合分布中采样——先根据输出概率选择一种情绪类别，再从相应的高斯分布中采样；也可以使用混合分布的均值，该均值是各情绪类别均值的加权平均，权重由输出概率决定。

表6.1：Ekman（1971年）的类别情绪对应的效价和唤醒度均值及标准差，数据取自 Russell 和 Mehrabian（1977）的用户研究。

	效价		唤醒度	
情感	Mean	Std	Mean	Std
愤怒	-0.51	0.20	0.59	0.29
厌恶	-0.60	0.20	0.35	0.41
fear	-0.64	0.20	0.60	0.32
joy	0.76	0.22	0.48	0.26
悲伤	-0.63	0.23	-0.27	0.34
惊讶	0.40	0.30	0.67	0.27

此外，我们引入一个参数，用于表示所有情绪类别中均值绝对值的最大值。利用该参数，我们计算出应用于所有效价-唤醒度分布的平移和缩放参数。通过这种方法，用户可以调整音乐生成器的条件化，从而选择更宽或更窄的情绪范围。以下公式展示了如何重新调整效价和唤醒度值的范围。

$$
\begin{aligned}
\mathbf{v}'_{\text{means}} &= \frac{2r \cdot (\mathbf{v}_{\text{means}} - \min(\mathbf{v}_{\text{means}}))}{\max(\mathbf{v}_{\text{means}}) - \min(\mathbf{v}_{\text{means}})} - r \\
\mathbf{v}'_{\text{stds}} &= \frac{2r \cdot \mathbf{v}_{\text{stds}}}{\max(\mathbf{v}_{\text{means}}) - \min(\mathbf{v}_{\text{means}})} \\
\mathbf{a}'_{\text{means}} &= \frac{2r \cdot (\mathbf{a}_{\text{means}} - \min(\mathbf{a}_{\text{means}}))}{\max(\mathbf{a}_{\text{means}}) - \min(\mathbf{a}_{\text{means}})} - r \\
\mathbf{a}'_{\text{stds}} &= \frac{2r \cdot \mathbf{a}_{\text{stds}}}{\max(\mathbf{a}_{\text{means}}) - \min(\mathbf{a}_{\text{means}})}
\end{aligned}
$$

我们选取 r 来表示所有情感类别中均值的新的最大绝对值，因此，总体范围变为 $2r$。上标撇号表示新值。粗体字母表示包含所有情感类别效价和唤醒度均值与标准差值的向量。在平移和缩放均值时，我们只缩放标准差，因为标准差仅表示数据的离散程度，而不表示数据的位置。最小值和最大值可直接从表6.1 中找到，如下：

$$
\min(\mathbf{v}_{\text{means}}) = -0.64, \quad \max(\mathbf{v}_{\text{means}}) = 0.76, \quad \min(\mathbf{a}_{\text{means}}) = -0.27, \quad \max(\mathbf{a}_{\text{means}}) = 0.67.
$$

6.2 实现细节

基于视频的音乐生成流程如图 6.1所示。在左侧分支中，我们进行情绪条件控制。首先，将所有

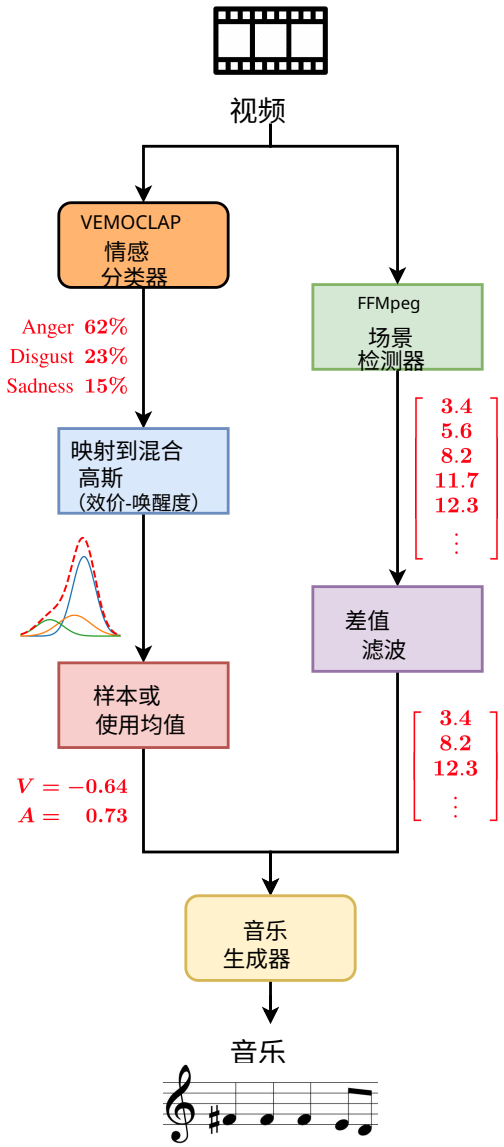


图6.1：我们的基于视频的音乐生成流程。箭头旁的值为样本值。

情感类别的均值的最大绝对值设为 0.8，并相应调整效价-唤醒度分布。VEMO- CLAP 模型对视频中的情绪进行分类，并输出一个概率分布。随后，将该分布映射为高斯混合，其中概率分布作为各情绪类别效价-唤醒度分布的权重。基于此，我们可以从该分布中采样，或使用混合分布的均值。为确保结果更加稳健，我们选择使用混合分布的均值。最终输出由效价和唤醒度两个值组成，它们将作为音乐生成器的情感条件。

右侧分支专注于时间条件控制。在这里，FFMpeg 场景检测器通过检测连续帧之间的平均像素差异是否

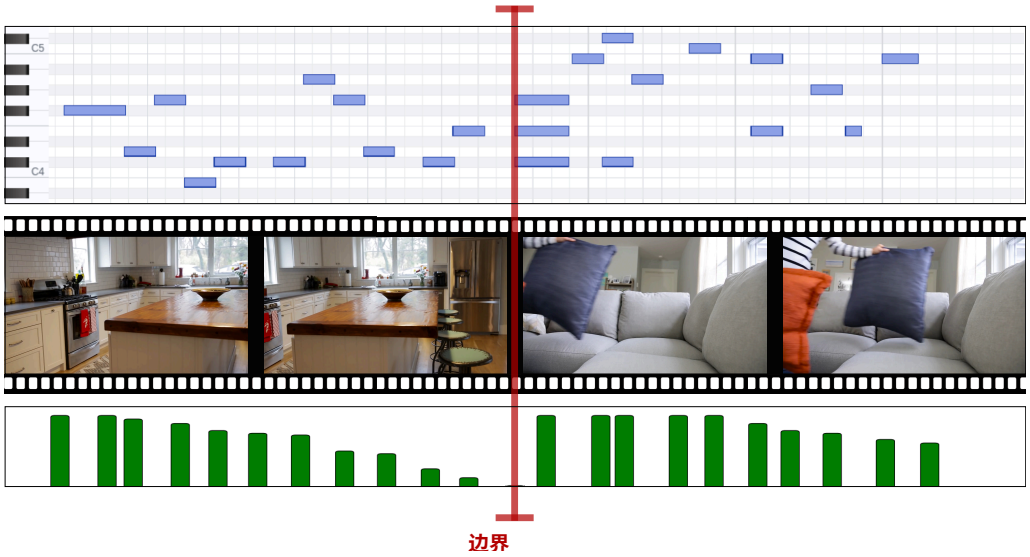


图 6.2：边界的图示说明。

超过像素值范围（0-255）的 27%，来识别场景切换位置。随后应用差分滤波器，移除彼此过近的位置，以免强和弦重叠得太过密集，形成刺耳的杂音。在我们的实验中，我们会丢弃间隔不足 4 秒的场景。由此得到的时间戳将作为音乐生成器的时间条件。

音乐生成器持续生成词元，直到生成音乐匹配视频的时值。将 MIDI 合成为音频波形后，会施加 3 秒淡出，以避免突然结束。在训练视频情绪分类器与音乐生成器——及其时间与情绪条件机制——之后，我们不对组合模型进行任何额外训练或微调。所有超参数均通过试错以及对输出的人工检查来确定。

图 6.2 图示说明了我们如何将音乐边界（强和弦）与视频边界（场景切换）对齐。在训练边界条件机制时，我们使用长时值和弦的时间戳作为目标边界。在基于视频的推理中，这些目标边界被替换为从视频中提取的场景切换时间戳。

在最上方一行，我们展示一段符号音乐，其中由三个或更多同时发音的音符组成、且时值超过设定阈值两拍的和弦定义一个音乐边界。这些音乐边界仅在训练期间使用，如第4.1节所述。中间一行展示一段视频，其中场景切换作为视频边界，并在推理期间使用。最下方一行说明边界偏移量，它们根据输入边界计算得出。

需要注意的是，这些图示仅用于说明；在此示例中，除边界处外，偏移量与音乐并未完全对齐。此外，在基于视频的推理过程中，我们并不期望音乐边界与视频边界完全匹配。如前所述，我们的音乐生成器会在输入边界附近生成强和弦，以确保节奏与其余生成音乐保持连贯，同时不打断

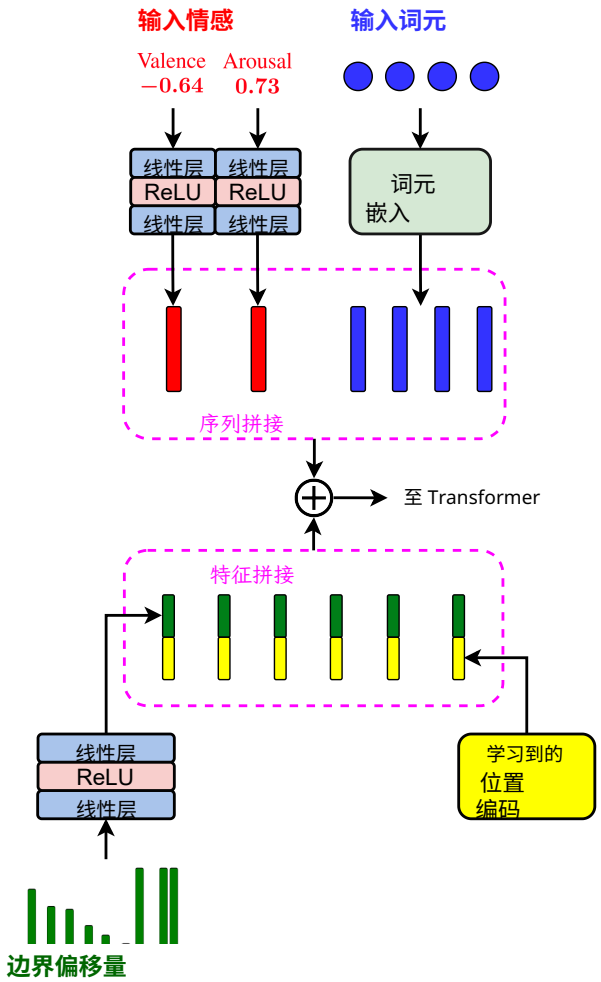


图 6.3：我们的音乐生成器。效价和唤醒度下方的数字是样本值。

其流畅性。

对于基于情感的 MIDI 生成，我们使用第4.2节所述的连续令牌模型。我们选择该模型，是因为它采用连续值条件，避免了离散化所带来的信息损失。此外，虽然其性能几乎与表现最佳的连续拼接模型相当，但不会增加模型的维度。不过，我们所有的情绪条件模块都兼容于总体模型的集成。

图6.3 展示了总体条件音乐生成流程，同时整合了情感条件控制与时间条件控制。情感条件控制通过沿序列维度将条件向量与词元嵌入进行拼接来实现。如第4.1节所述，时间边界条件控制通过沿特征维度将边界偏移量与位置编码进行拼接来实现。
Transformer的输入表示为：

$$\mathbf{X}_{\text{trans}} = \text{拼接}_s(\text{FFN}_v(x_v), \text{FFN}_a(x_a), \text{嵌入}(\mathbf{x}_t)) + \text{拼接}_f(\text{FFN}_b(\mathbf{b}), \mathbf{W}^{\text{pe}}) \quad (6.1)$$

其中x_v 和x_a 分别为效价和唤醒度输入，$\mathbf{x}_t$ 为输入词元序列，$\mathbf{b}$ 表示输入边界偏移，$\mathbf{W}_{\mathbf{pe}}$ 为学习式位置编码。前馈网络FFN_v、FFN_a和FFN_b 分别处理效价、唤醒度和边界偏移输入。运算Concat_s 和Concat_f 分别表示沿序列维度和特征维度进行拼接。这两种条件机制旨在易于集成且互不干扰。此外，这两种条件机制均不修改Transformer的身体，因此只需移除或替换条件模块，即可轻松微调我们已训练的模型，以用于其他类型的音乐生成任务。

我们选择了音乐生成器的规模，以确保能够与后文详细介绍的最先进方法进行公平比较。我们的音乐生成器由 11 层、8 个注意力头组成，维度为 512，共包含 3700 万个参数。训练时，上下文长度为 1216 个词元，批量大小为 64。我们采用交叉熵损失进行训练，前 30 万步的学习率为 2e-4，接下来的 30 万步则为 5e-5。我们使用 Adam优化器，并将梯度裁剪至范数 1（Kingma & Ba，2015年）。我们不采用任何正则化方法，也未观察到过拟合，因为模型是在大规模密集标注数据上训练的，并且会预测输入序列中的每个词元。

6.3 评估

我们对模型进行了客观评估和主观评估，并将其与两种用于从任意视频生成 MIDI 的开源最先进方法进行比较：Video2Music（Kang 等人，2024）和 CMT（Di等人，2021），二者均已在第2.6节介绍。

Video2Music 与我们的方法有相似之处，二者都利用低层视频特征和高层情感条件。不过，音乐生成过程存在一项重大差异：他们的模型只生成和弦序列，随后再通过固定的手工琶音模式将其转换为钢琴旋律。这些模式都依赖粗粒度时间网格，因此难以捕捉人类演奏中富有表现力的时序。相比之下，我们的模型以直接且通过学习实现的方式生成多乐器音乐，并采用具有精细时间精度的事件表征。此外，Video2Music 需要用户提供调性和初始和弦，而我们的方法完全自动化。

可控音乐 Transformer（CMT）与我们的方法类似，能够直接生成多乐器音乐，并在 Lakh 钢琴卷帘数据集上训练。不过，与我们的模型不同，它仅依赖时序、运动显著性和运动速度等低层视频特征，不包含情感等高层特征。此外，与 Video2Music 类似，它将音符位置限制在节拍细分上，因此难以捕捉表现性节奏中的细微差异。

在最终评估中，为了将我们的结果与 Video2Music 和 CMT 进行比较，我们将模型规模设定为与它们相当：我们的模型有 3700 万个参数，而 CMT 有 3900 万个，Video2Music 有 3300 万个。对于所有方法，我们都使用 Fluidsynth 将输出 MIDI 文件合成为音频波形，并进行峰值归一化，最高至 -3 dB。

6.3.1 数据集

我们在 Lakh Pianoroll-5 数据集上训练音乐生成器（Dong 等人，2018年）。在客观和主观评估中，我们首先使用 EmoMV-C 数据集。该数据集由音乐视频组成，其中包含用作客观评估真值的音乐音轨（Thao 等人，2023年）。为避免数据泄漏，我们在情绪分类时排除听觉特征。为了评估模型在不同视频类型上的泛化能力，我们还纳入了匹兹堡广告（Ads）数据集，其中包含可能有音乐、也可能没有音乐的广告视频（Hussain 等人，2017年）。广告通常依靠音乐来最大程度地提高观众参与度，因此 Ads 数据集提供了一个真实场景来进行评估（Zander，2006）。

我们使用完整的 EmoMV-C 验证集，其中包含 48 个时长为 30 秒的视频。我们对 Ads 数据集进行筛选，使其总时长与 EmoMV-C 相匹配。我们选取与音乐情绪分类中常用的四种基本情绪相关的视频（Y. Yang 等，2008），覆盖 Russell 的效价—唤醒模型（Russell，1980）中的各个象限： 愉快、 平静、 愤怒和 悲伤，从而为评估提供广泛的情绪空间覆盖。虽然所选情绪中的一些与 Ekman 的类别（Ekman，1971年）有所重叠，但我们在选择评估视频时避免明确使用 Ekman 的情绪。这样可以避免评估偏向我们的方法，因为我们的视频情绪分类器经过训练，输出的是 Ekman 的类别。我们排除时长不足 1 分钟的视频。为确保选择不带偏见，我们使用由 YouTube 随机生成的 YouTube ID。我们按字母顺序对这些 ID 排序，并从每个情绪类别中选取前六个视频，共得到 24 个评估视频。最后，我们将视频剪辑至统一的 1 分钟时长。随后，将这些评估视频输入所比较的模型。

6.3.2 客观评价

在客观评价中，我们首先使用 EmoMV-C 数据集中的真实值音频来评估音乐相似度。尽管真值并非 MIDI 格式，但所有方法都会生成 MIDI 输出，并将其合成为音频，从而能够在音频领域进行公平比较。我们使用预训练的音乐标签模型（Won et al., 2020年）计算生成音乐与真实值音乐标签之间的 Kullback-Leibler 散度（KL）。这衡量了生成音乐中的语义标签分布与真实值的分布有多接近。我们不采用常用的 Fréchet 音频距离，因为它比较的是适用于无条件生成的总体音频分布（Kilgour et al., 2019年），而基于视频的音乐生成需要逐样本评估，以衡量生成音乐与每段视频独特内容的对齐程度。

此外，我们还使用 EmoMV-C 和 Ads 数据集评估时序音视频对齐（AV-Align）分数，该分数是视频峰值与音频峰值之间的交并比（Yariv 等，2024）。视频峰值根据光流幅值的均值检测得出（Horn & Schunck，1981年），而音频峰值则对应音频起始点（Böck & Widmer，2013年）。我们在两种时间分辨率（窗口）下计算 AV-Align 分数：一帧视频画面（30 FPS 时为 33.3 毫秒）和一秒。对于每个窗口，我们检测视频和音频中的局部峰值，并

在相同时间范围内进行匹配，以评估对齐情况。最后，我们报告各模型为一分钟视频生成音乐时的平均运行时间，其中包括模型初始化。

6.3.3 主观评估

使用 EmoMV-C 和广告数据集中的评估视频，我们通过三个模型生成配乐，并创建调查页面，每页包含两个视频。对于每个视频，比较方法生成的三个版本的音乐并排展示，并以匿名方法名称标注；每种方法的输出在左侧、中央和右侧位置出现的次数相同。

我们招募了 153 名远程参与者，并请他们依据以往研究采用的标准（Di 等人，2021；Kang 等人，2024）进行排名：

- 音乐丰富度（*MR*）——音乐的丰富程度和多样性，与视频内容无关。
- 音乐质量（*MQ*）——音乐的总体质量，与视频内容无关。
- 情感匹配（*EM*）——音乐在情感方面与视频的匹配程度。
- 时序匹配（*TM*）——音乐在节奏和时序方面与视频的同步程度。
- 整体匹配（*OM*）——音乐与视频在整体上的匹配程度。

每个模型都被赋予唯一排名：1（最佳）、2 或 3（最差）。图 6.4 展示了用户调查的一个样本视图。模型名称被匿名化为方法 A、方法 B 和方法 C。用户需将方法名称从左侧拖到右侧，从而对方法进行排名，以确保漏答某个问题不会导致错误。为方便读者，我们还在线提供了八个视频的输出样本，其中方法名称已去匿名化，且每种方法的输出均出现在相同位置，以便清晰对照。[1]

6.4 结果与讨论

我们使用平均排名，在所有主观标准上比较 CMT（Di等人，2021）、Video2Music（Kang 等人，2024）以及我们的方法 EMSYNC。表6.2、表 6.3 和表 6.4 分别给出了第 1 组、第 2 组以及全体参与者的结果。排名先按每个视频取平均，再在所有视频上取平均。分数越低表示排名越好（1 为最佳，3 为最差）。各视频间的标准差显示在括号中。

图 6.5、6.6和6.7 分别展示了第 1 组、第 2 组和所有参与者的参与者回答频率堆叠条形图。每个图表按评估指标和方法分组，显示所分配的每个排名的频率。

[1] https://serkansulun.com/emsync

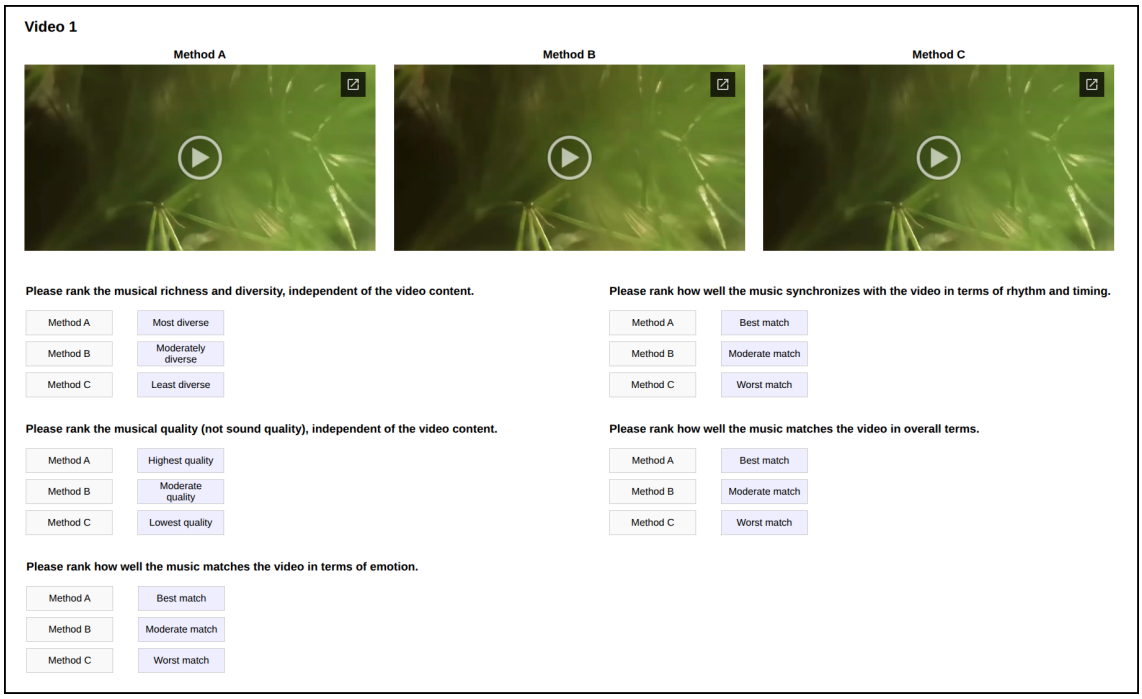


图 6.4：我们的基于视频的音乐生成调查截图。

表 6.2：具有自我报告的音乐理论理解的参与者的结果。较低的值表示更好的排名（1 =最佳，3 = 最差）。值以均值（标准差）报告。

	音乐丰富度	音乐质量	情感匹配	时序匹配	总体匹配
CMT (Di等人，2021)	1.85 (0.56)	2.44 (0.60)	2.00 (0.53)	2.12 (0.76)	2.08 (0.64)
Video2Music (Kang 等人，2024)	2.52 (0.71)	1.85 (0.67)	2.12 (0.86)	1.96 (0.76)	2.17 (0.83)
EMSYNC（我们的方法）	<风格 id='1'>1.62 </风格>(0.71)	**1.71** (0.76)	**1.88** (0.85)	**1.92** (0.75)	**1.75** (0.79)

表 6.3：没有自我报告的音乐理论理解的参与者的结果。较低的值表示排名更好（1 =最佳，3 =最差）。值以均值（标准差）表示。

	音乐丰富度	音乐质量	情感匹配	时序匹配	总体匹配
CMT (Di等人，2021)	1.79 (0.64)	2.56 (0.61)	2.46 (0.57)	2.31 (0.75)	2.35 (0.58)
Video2Music (Kang 等人，2024)	2.71 (0.36)	1.77 (0.53)	1.85 (0.73)	1.98 (0.65)	2.08 (0.73)
EMSYNC (我们的方法)	**1.50** (0.49)	**1.67** (0.67)	**1.69** (0.64)	**1.71** (0.69)	**1.56** (0.66)

表 6.4：所有参与者的结果。较低的值表示更好的排名（1 = 最佳， 3 = 最差）。值以均值（标准差）报告。

	音乐 丰富度	音乐 质量	情感 匹配	时序 匹配	总体 匹配
CMT (Di等人，2021)	1.82 (0.60)	2.50 (0.60)	2.23 (0.59)	2.22 (0.75)	2.22 (0.62)
Video2Music (Kang 等人，2024)	2.61 (0.57)	1.81 (0.60)	1.99 (0.80)	1.97 (0.70)	2.12 (0.78)
EMSYNC （我们 的方法）	**1.56** (0.61)	**1.69** (0.71)	**1.78** (0.75)	**1.81** (0.72)	**1.66** (0.73)

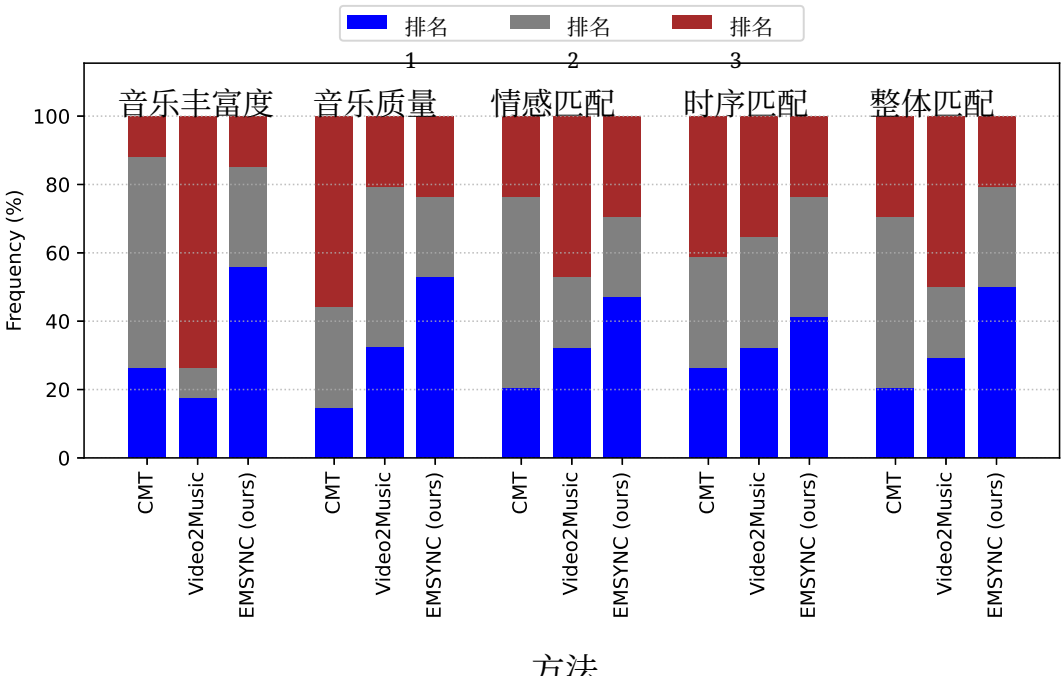


图 6.5：具有自我报告的音乐理论理解的参与者的堆叠条形图结果。

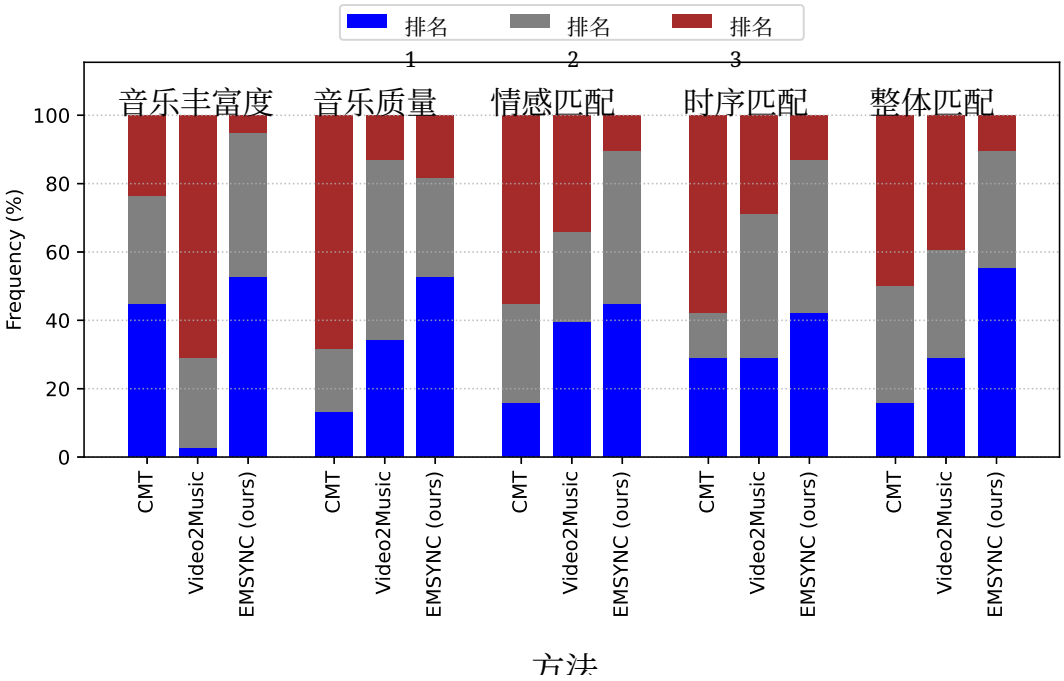


图 6.6：没有自我报告的音乐理论理解的参与者的堆叠条形图结果。

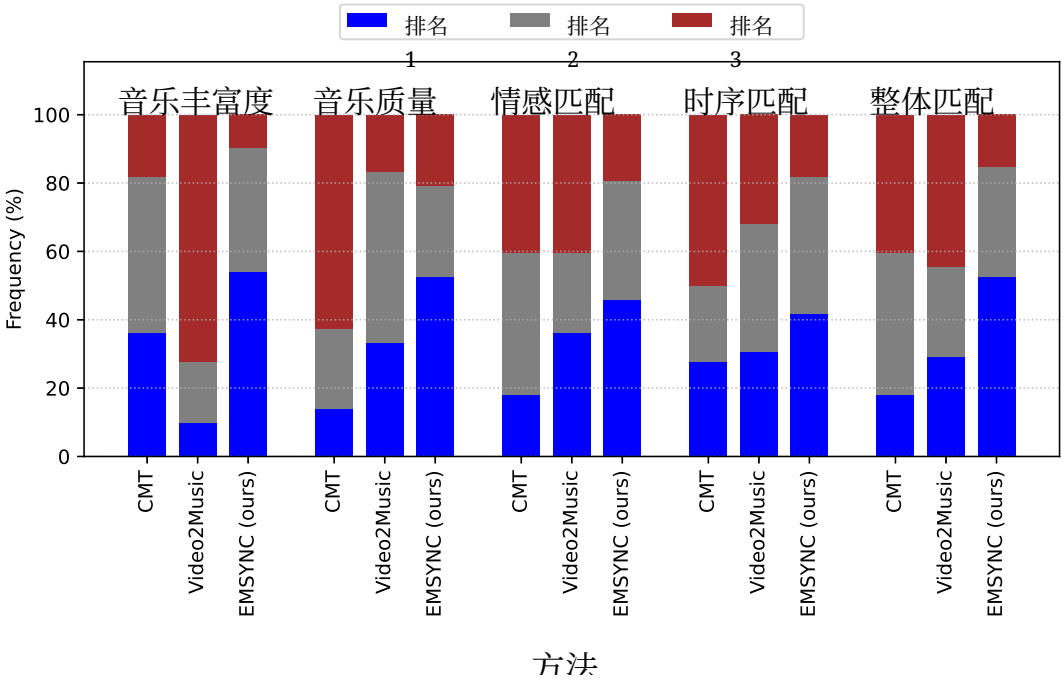


图 6.7：所有参与者的堆叠条形图结果。

最后，我们报告各模型为一分钟视频生成音乐时的平均运行时间（包括模型初始化）：EMSYNC 为 1.42 分钟，CMT 为 3.55 分钟，Video2Music 为 1.61 分钟。

在音乐丰富度方面，Video2Music 在两组参与者中的平均排名都是最差的。这可能归因于它依赖固定的琶音模式，或仅限于生成仅钢琴输出。具备音乐理论知识的参与者较少将 CMT 排在第一。根据我们对模型输出的初步审查，一个可能的解释是，CMT 会利用视频运动等稀疏约束来调整音符密度，这可能导致节奏不连贯的结果。没有音乐理论知识的参与者可能会将此视为增加了变化，并提升了音乐丰富度，而其他人则未必如此。此外，没有音乐理论知识的参与者很少将 Video2Music 排在第一，这可能是因为即使不具备任何音乐理论知识，也很容易看出其输出由于局限于单一乐器而缺乏变化，而其他方法则会生成多乐器音乐。

在音乐质量方面，CMT 在两组参与者中的平均排名都是最差的。这可能再次是由于它持续调制音符密度。在情感匹配方面，CMT 在两组中被评为最佳的频率都最低，这符合预期，因为它仅依赖低层特征，缺乏诸如情感之类的高层条件控制。然而，一个令人惊讶的观察是，具备音乐理论知识的参与者也最不可能将 CMT 评为最差。因此，它在这一类别中的平均排名优于 Video2Music。这或许凸显了高层视频情感分类本身的难度，并表明我们的模型和 Video2Music 的情感分析组件仍有改进空间。

在时序匹配方面，我们的方法尤其受到不具备音乐理论知识的参与者青睐。这或许支持了我们的假设：普通大众更容易感知与场景切换等稀疏时间线索的对齐。在整体匹配方面，我们的方法效果更加显著。它在所有参与者群体中都取得了最佳表现，而其他方法的平均排名均低于 2。

在所有指标和所有组别上，EMSYNC 的表现始终优于其他方法。这些结果证实，EMSYNC 生成的音乐最多样、质量最高，在情感对齐、时间同步和整体视频兼容性方面也表现最佳，同时还是计算效率最高的模型。

本章总结了我们对完整基于视频的音乐生成系统的介绍。读者或许还记得我们最初在图1.1中概述的模型流程；值得注意的是，除了音频生成器之外，我们现已详细讨论了每个组成部分。为了构建一个能够生成可聆听输出的工作原型，我们使用了预设声音库和 Fluidsynth 软件，由它将这些声音映射到各个 MIDI 音符。这种方法的主要局限是音色多样性不足：由于声音库是固定的，生成的音频缺乏细腻度和真实感，与人类音乐家的演奏相比，往往听起来过于机械。

虽然已有多种方法能直接在音频领域生成音乐，但它们通常

仅以文本为条件，并以黑箱模型的方式运行（Copet 等人，2023年；Evans 等人，2024；Lam 等人，2023年；H. Liu 等人，2024）。由于其输出纯粹以音频形式存在，因此无法编辑或解释。因此，与其称之为音乐作品，不如将其称为音乐声音，这或许更为准确。这种黑箱方法与人类创作音乐的过程形成对比；人类创作通常分为两个不同阶段：作曲（确定和弦、音符、旋律和结构），随后是制作（通过演奏生成音频）。我们认为，自动音乐生成的一个有前景的方向，是在这两个阶段分别采用深度神经网络：通过 MIDI 生成进行作曲，通过 MIDI 条件音频生成进行制作。尽管基于 MIDI 的音频生成研究仍在进行，但目前的工作往往局限于单轨钢琴音乐，这主要是由于成对的多乐器 MIDI 与音频数据集稀缺（Hawthorne 等人，2019年；Sambaragi 等人，2024）。下一章将对音频生成及其挑战进行初步分析，尤其探讨如何依靠合成数据弥补真实世界音频录音的不足。

第7章

音频带宽扩展

由于我们的基于视频的音乐生成器会以符号音乐格式创作配乐，因此需要借助计算机软件和音色库将其合成为音频。音色库为特定乐器的特定音高提供固定音色。然而，与人类音乐家演奏的音乐相比，最终效果非常机械、不真实。因此，我们研究音频增强，以迈出改善符号音乐合成音频质量的第一步。

目前还没有包含配对音频（由人类演奏）和 MIDI 的多乐器音乐数据集，而这类数据集是训练端到端模型所必需的；该模型应能根据合成的 MIDI 生成听起来像人类演奏的音频。即使有这样的数据集，从 MIDI 创作出类似人类演奏的音频仍是一个定义不明确的问题，因为同一首音乐作品（即 MIDI 文件）可以衍生出无穷多种人类演奏方式。由于缺乏包含 MIDI—音频配对数据的多乐器数据集，一种可行的方法是利用音色库从 MIDI 生成合成音频，并以此训练深度神经网络。然而，这些训练所得模型的输出质量会受合成音频质量的限制，导致深度神经网络无法发挥作用。

我们研究了另一个定义不明确的音频增强问题，即音频带宽扩展。这项任务旨在根据带限音频创建全频带音频。由于有限频带中的信息缺失，而这些缺失信息可能有多种候选，因此这项任务定义不明确。我们还研究了这个问题，因为它涉及合成数据。总体而言，研究音频带宽扩展能带来宝贵的见解，这些见解也适用于其他利用合成数据的定义不明确的音频生成任务，例如 MIDI到音频生成。研究音频带宽扩展的一个关键优势是可用数据十分丰富：只需对任意全带宽音频应用各种低通滤波器，就能生成无限量的输入数据。

在考虑通过音频带宽扩展来增强真实世界录音时，并不存在全带宽版本，因此也没有可作为深度神经网络目标的“真实值”。为此，训练数据通常通过对全带宽录音进行低通滤波来获得。然而，由于真实世界的带限样本并非由某种假想的通用

数字低通滤波器处理而成，因此，开发依赖于对带宽缩减过程进行粗略近似的稳健带宽扩展技术颇具挑战，而这也使其难以泛化至未见过的录音。在使用某一种类型的低通滤波器生成的训练数据上，训练后的深度神经网络可能表现良好，但面对经过不同类型低通滤波器处理的音频内容时，却可能无法泛化。在我们的工作中，我们将这种现象称为 滤波器过拟合 ，可将其理解为滤波器泛化能力不足。

尽管在现有使用深度神经网络进行音频带宽扩展的研究中，低通滤波十分普遍，但据我们所知，迄今尚无研究对滤波器泛化这一主题进行过全面考察。我们认为，无法泛化到各种类型的信号劣化，是创建可部署于真实世界的音频增强模型时面临的一项重要挑战。在本研究中，我们对滤波器泛化进行了严格分析，使用两种广受欢迎的深度神经网络架构，评估模型在复杂音乐信号带宽扩展任务中，对用于预处理输入数据的不同滤波器的泛化能力。

为了评估样本过拟合，我们使用与训练数据集划分不同的测试数据划分；对于训练好的模型而言，这些数据属于未见数据。此外，为了评估滤波器泛化，我们使用与预处理训练输入数据的滤波器不匹配的滤波器，对测试输入数据进行预处理，即使用未见过滤器 ，并将其与测试数据和训练数据的预处理滤波器相匹配的测试设置进行比较，即已见滤波器。我们认为，使用未见过滤器进行测试可视为对真实世界信号退化的模拟，因为其真正的退化函数是未知的。

我们评估了文献中用于提升泛化能力的两种不同正则化方法。具体而言，我们比较了使用数据增强、批归一化和随机失活，与不使用任何正则化方法的基线。我们提出了一种新颖的数据增强技术：使用一组不同的低通滤波器对输入数据进行预处理，其中从不使用未见滤波器处理测试数据。我们通过在训练迭代过程中跟踪模型的演奏表现，并使用已见滤波器和未见滤波器进行验证，来考察训练过程。

我们使用的深度神经网络模型之一是*U-Net*，它最初用于生物医学图像分割（Ronneberger 等，2015年），后来用于歌声分离等音频信号处理任务（Jansson 等，2017年），最终也用于音频增强（S. Kim 和 Sathe，2019年；Kuleshov 等，2017年；T. Lim 等，2018年；Macartney 和 Weyde，2018年）。除了 U-Net，我们还使用深度残差网络模型（*ResNet*）（He 等，2016），因为它是信号处理任务中应用最广泛的深度神经网络架构之一。尽管 U-Net 是近期音频处理文献中颇受欢迎的架构，但据我们所知，在音频处理领域尚无研究将 U-Net 与成熟的 ResNet 基线进行比较。图像处理和医学影像领域中存在少量比较研究，但这些研究要么没有说明所比较模型的参数数量（Rempfler 等，2017年），要么 ResNet 的参数数量显著少于 U-Net（Chiou 等，2018年；Ghodrati 等，2019年；Wu 等，2018年）。在所有这些研究中，ResNet 的表现都略胜于 U-Net。我们还展示了 U-Net 与

ResNet 的比较，其中两者的参数数量相近。

我们的主要发现表明，U-Net 和 ResNet 都会出现滤波器过拟合，但程度不同；与较为传统的正则化技术相比，采用多滤波器数据增强是缓解这一过拟合问题、进而改善带宽扩展中滤波器泛化的一种颇具前景的方法。

接下来，我们将详细介绍所采用的方法，首先介绍所使用的模型 y.

7.1 模型

本节将定义两种基线模型：U-Net 和 ResNet。对于这两种模型，我们都采用 Kuleshov 等人（2017年）的方法，使用原始音频作为输入，而不是使用时频变换（例如 Miron 和 Davies（2018年）的方法）。因此，无需在输出中进行相位重建。不过，由于我们研究的是带宽扩展，而非音频超分辨率本身，因此输入不会经过下采样。由此，我们所有模型的输入和输出尺寸都相同。

7.1.1 U-Net

U-Net 架构（Ronneberger 等，2015年）与其前身自编码器类似，由两个主要部分组成。第一部分包含下采样层，第二部分则包含上采样层，如图 7.1（左）所示。在 U-Net 中，同一尺度上的各个下采样层和上采样层通过堆叠连接相连，例如，第一个下采样卷积块的输出会与最后一个上采样卷积块的输入堆叠。与增强模型中的常见做法一样，网络还采用了从输入到输出的加性连接，因此网络只需建模输入与目标信号之间的差异，而不必从头生成目标信号。

在下采样部分，采用步幅为 2 的一维卷积层，从而有效地将激活长度减半。借用图像处理领域的术语，上采样部分包含“子像素”层（也称为像素重排器）（Shi 等，2016），可将激活长度加倍。子像素层通过将取自不同通道的样本交织到空间维度中来工作，从而有效地使通道长度减半。参数数量的设置旨在复现使用 U-Net 进行音频超分辨率的原始工作（Kuleshov，2020年；Kuleshov 等，2017年），本文将其统称为 *Audio-SR-U-Net* 。最终得到的网络包含 56.4 百万个参数。

7.1.2 ResNet

使用许多层训练原始前馈神经网络时，一个常见问题是“梯度消失”问题：由于反复相乘，反向传播到最早几层的梯度会趋近于零。残差网络（He 等人，2016）通过使用残差块消除了这一问题，这些残差块只对输入之间差异的一部分进行建模

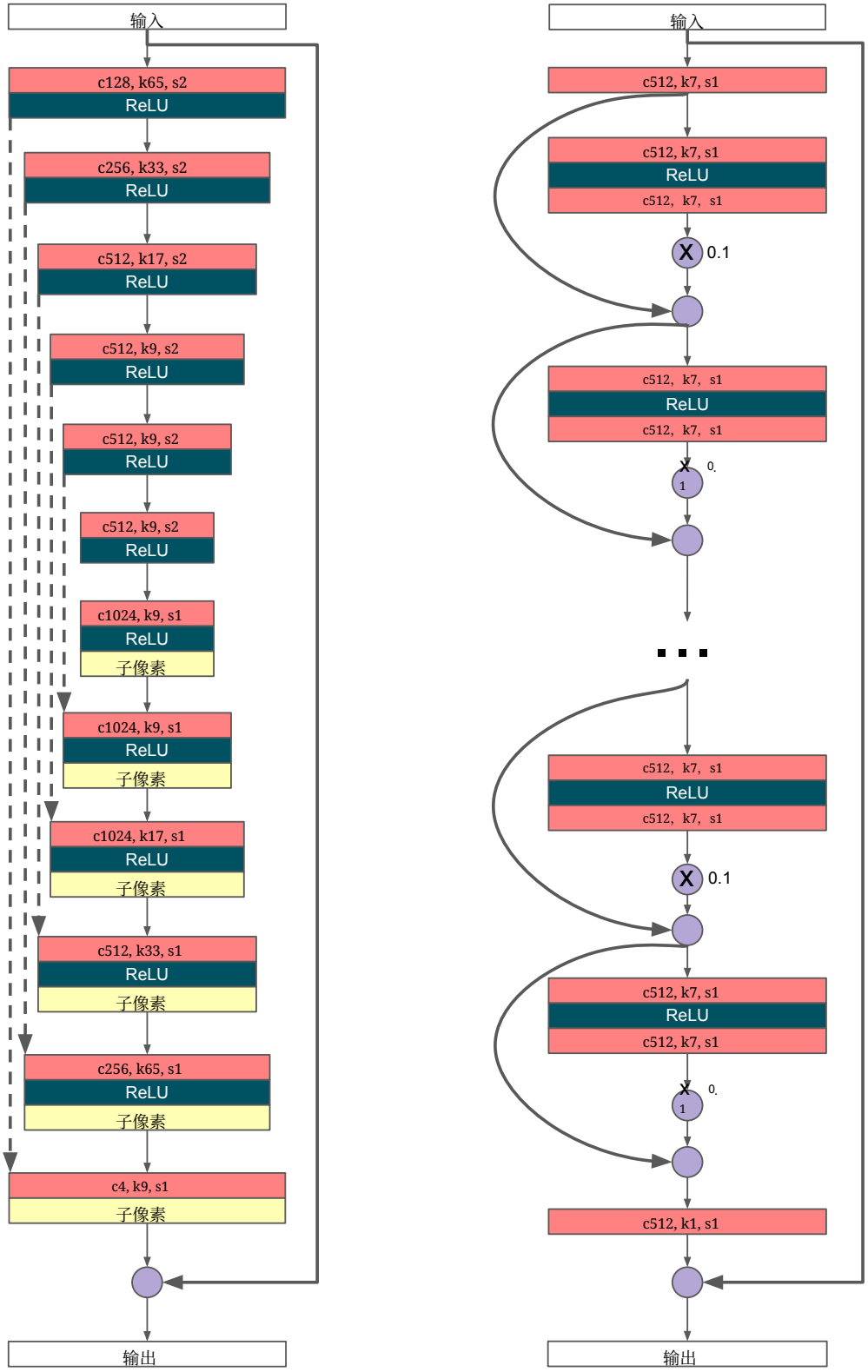


图 7.1：所使用的音频带宽扩展模型。（左）U-Net 模型，虚线表示堆叠连接。（右）包含 15 个残差块的 ResNet 模型。c、k 和 s 分别表示卷积层的通道大小、卷积核大小和步幅。

和输出。通常，每个残差块都包含两个卷积层，并在它们之间加入一个非线性函数。非常深的模型会采用残差缩放，即将每个残差块的输出乘以一个较小的数，例如 0。1，然后与其输入相加，以进一步稳定训练。我们的 ResNet 模型如图7.1 （右）所示。

与 U-Net 不同，ResNet 的激活长度在整个网络中保持不变。这样我们就能避免时间信息的任何损失，因为我们的目标是生成与输入等长的高分辨率输出。请注意，我们采用了简单的设计，除最后一层外，所有卷积层的参数数量都相同。与 U-Net 的实现类似，该模型共有 55。1 百万个参数。

在所有模型中，所有卷积都会应用适当的零填充，以保持激活尺寸不变。下采样卷积也不例外，因为下采样效果是通过步幅卷积实现的。

为了分析泛化能力，我们进行了消融研究，其中纳入了常见的防止过拟合的方法，这些方法称为正则化方法。

7.2 正则化方法

现在，我们将详细介绍标准的正则化方法，例如随机失活和批归一化，以及我们新颖的数据增强策略。

7.2.1 随机失活

防止过拟合的最简单、最常见的方法之一是随机失活，即按固定概率丢弃激活单元（Srivastava 等人，2014）。这会在隐藏层中引入噪声，并防止神经元过度协同适应。

尽管在文学中，随机失活很大程度上已被批归一化取代，尤其是在残差网络中，但新一代最先进的残差模型——即宽残差网络（Zagoruyko 和 Komodakis，2016）仍使用随机失活。此外，*Audio-SR-U-Net*的开源实现[1] 使用随机失活层而非批归一化，因此我们在 U-Net 模型中沿用了这一实现，并在每个上采样卷积层后使用随机失活层。在 ResNet 模型中，我们将随机失活层放置在每个残差块的两个卷积层之间。所有实验的丢弃率均设为 0。5。

7.2.2 批归一化

训练深度神经网络时，更新模型参数会有效改变下一层输入的分布。这被定义为内部协变量偏移，而批归一化通过对层输入进行归一化来解决这一问题（Ioffe 和 Szegedy，2015年）。尽管批归一化主要是为加速训练而提出的，但它也具有正则化作用。由于归一化参数是基于每个批次学习得到的，因此它们只能

[1] https://github.com/kuleshov/audio-super-res

提供对真实均值和方差的含噪估计。使用这些估计参数进行归一化会在隐藏层中引入噪声，并减少过拟合。

对于 U-Net，我们采用了*Audio-SR-U-Net* 模型（Kuleshov 等，2017年），并在每个降采样卷积层之后插入批归一化层。对于 ResNet，则在每个卷积层之后使用批归一化。

7.2.3 数据增强

为了提高深度神经网络的样本泛化能力，会使用数据增强：在输入数据样本送入深度神经网络之前对其进行变换，从而有效增加训练样本的数量和多样性。数据增强在基于图像的任务中十分常见，主要利用旋转、翻转或裁剪等几何变换（Taylor 和 Nitschke，2018年）。将这类几何变换应用于音乐信号时，通常无法产生逼真的样本。尽管已有一些针对音乐信号数据增强的研究（McFee 等，2015年），但这些研究主要通过时间拉伸和移调等有损变换，提高分类任务（例如乐器识别）的鲁棒性，而这些变换不适用于此处的情境。

由于我们的主要目标是探索并改进滤波器泛化能力，因此我们提出了一种数据增强方法，在训练过程中使用多种不同类型的滤波器。未采用数据增强的基线方法使用单滤波器训练设置，具体而言，使用六阶第一类切比雪夫滤波器，记作“Chebyshev-1, 6”。而在采用数据增强时，我们使用多滤波器训练设置，在训练期间针对每个输入样本随机选取由八种不同滤波器组成的集合。这八种滤波器包括不同阶数的第一类切比雪夫、贝塞尔和椭圆滤波器。为评估滤波器泛化能力，我们将六阶巴特沃斯滤波器保留为未见滤波器。表7.1总结了这些滤波器，其在评估期间的使用情况详见第7.4节。图7.2展示了这些滤波器不同频率幅度响应的图形概览。

将三种正则化方法与两种基线架构结合后，我们得到以下八种用于实验的模型：基线 U-Net（*U-Net*）、采用数据增强的 U-Net（*U-Net DA*）、采用批归一化的 U-Net（*U-NetBN*）、采用随机失活的 U-Net（*U-Net DO*）、基线残差网络（*ResNet*）、采用数据增强的残差网络（*ResNet DA*）、采用批归一化的残差网络（*ResNet BN*）以及采用随机失活的残差网络（*ResNet DO*）。

7.3 数据集

带宽扩展的机器学习方法通过使用数据集来构建问题，这些数据集包含每个音频信号的全带宽（高质量）版本和带限（低质量）版本。构建这类样本对的一种直接方法是获取高质量数据集，然后对其进行低通滤波。尽管音乐音频数据集有很多，尤其是在音乐信息检索领域，但其中许多数据集汇集自不同来源

表 7.1：在两种不同的训练设置下使用的低通滤波器类型和阶数：单滤波器 （无数据增强）和多滤波器 （数据增强）。

	单滤波器 （无数据增强）	多滤波器 （数据增强）
训练 验证，使用已见滤波器	六阶第一类切比雪夫滤波器	六阶第一类切比雪夫滤波器 切比雪夫一型，8 切比雪夫一型，10 切比雪夫一型，12 贝塞尔，6 贝塞尔，12 椭圆，6 椭圆，12
使用验证集进行。 未见过滤器 使用测试与未见过滤器	巴特沃斯 ，6	巴特沃斯 ，6
使用……进行测试 已见滤波器	六阶第一类切比雪夫滤波器	六阶第一类切比雪夫滤波器

（包括研究人员的个人音频收藏），且其中的音频内容通常经过压缩（例如采用有损 MP3/AAC 编码），因此并非严格意义上的全带宽音频。

除了要求音乐音频内容具有全带宽外，我们提出的方法旨在不受音乐风格限制。为此，任何未压缩的全带宽音乐内容都可用作训练材料；不过，为了确保结果可复现，我们选用了以下两个公开数据集，它们包含全带宽、立体声和多轨音乐音频：MedleyDB（2.0 版）（Bittner 等人，2016）和 DSD100（Liutkus 等人，2017年）。在每个数据集中，音频内容的采样率均为 44100 Hz（即 CD 音质），因此带宽为 22050 Hz。

MedleyDB 包含 121 首歌曲，而 DSD100 分为训练集和测试集，各包含 50 首歌曲。由于包含分离的多轨分轨，这两个数据集在音乐混音和音频源分离研究中都得到了广泛应用。然而，我们旨在研究多乐器音乐的带宽扩展，而非孤立的单一乐器，因此仅保留每首歌曲的立体声混音。为了创建带限输入样本，我们采用截止频率固定为 11025 Hz 的低通滤波器，即原始带宽的一半。

使用 DSD100 测试集进行测试，DSD100 训练集中的最后 8 首歌曲用于验证，其余所有 DSD100 训练集歌曲以及整个 MedleyDB 数据集则用于训练。

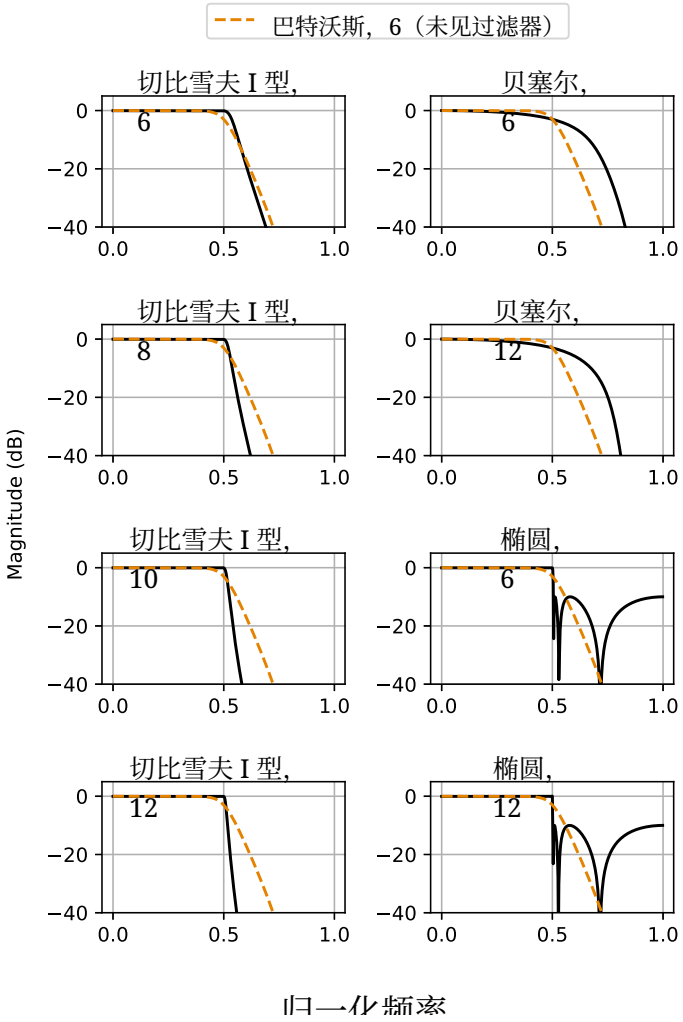


图 7.2：训练滤波器的频率响应。每张图中均叠加了未见滤波器（6 阶巴特沃斯滤波器）的频率响应。

7.4 评估

我们介绍用于验证的评估方法，它揭示训练过程中的过拟合趋势；还介绍用于测试的评估方法，它反映最终演奏表现，并对所用指标作出说明。

7.4.1 指标

由于我们的研究更关注泛化问题，而非演奏表现，因此我们采用一个成熟的指标——信噪比（SNR）来评估实验：

$$\mathrm{SNR}(x,\hat{x}) = 10\log_{10}\frac{||x||_2^2}{||x-\hat{x}||_2^2} \tag{7.1}$$

其中，x 是参考信号，$\hat{x}$ 是其近似值。计算二范数时，使用的是信号的立体声形式。就本研究的具体情境而言，我们认为信噪比适合用于考察过拟合，因为我们的模型采用均方损失进行训练，而最小化均方损失对应于最大化信噪比。

为了进一步了解演奏表现，我们使用 VGG 距离评估输出音频样本的感知质量。Y. Li 等人最近在音乐增强评估中采用了这一方法（2020年）。两个音频样本之间的 VGG 距离定义为它们的嵌入之间的距离，这些嵌入由在音频分类任务上预训练的 *VGGish* 网络生成（Hershey 等人，2017年）。Manocha 等人（2020年）表明，与语音质量感知评估（PESQ）（Rix 等人，2001）和虚拟语音质量客观听评器（ViSQOL）（Hines 等人，2012年）等手工设计指标相比，深度嵌入之间的距离与人工评估的相关性更高；这一结论适用于包括带宽扩展在内的各种音频增强任务。 *VGGish*嵌入也用于计算 Fréchet 音频距离（FAD）。这是一种最先进的评估方法，用于评估一组输出样本的感知质量（ Kilgour 等人，2019年）。然而，由于 FAD 用于比较两组样本，而非单个音频信号，因此不适用于我们的情况。

为了获得 VGG 嵌入，我们使用了*VGGish* 神经网络的开源实现[2]。除将采样频率设置为 44100 Hz、最高频率设置为 22050 Hz 外，我们使用了默认参数。与信噪比计算不同，参考实现在计算 VGG 嵌入之前会将立体声信号下混为单声道。经过后处理，嵌入的取值范围为 0 到 255。与 Manocha 等人（2020年）类似，我们采用平均绝对距离来定义 VGG 距离：

$$\mathrm{VGG}(x,\hat{x}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{7.2}$$

其中 x 为参考信号，$\hat{x}$ 为其近似值；y 与 $\hat{y}$ 分别为它们的嵌入。n 为嵌入张量的大小，取决于 x 的长度。

7.4.2 测试

为评估我们模型的总体性能，我们在训练结束时进行一次测试。DSD100数据集的测试集预留给我们的测试阶段。由于GPU内存限制，我们的神经网络无法在单次前向传播中处理完整长度的歌曲，因此它们处理不重叠的音频片段，随后将输出拼接起来，生成完整长度的输出歌曲。对于 VGG 距离和信噪比，我们首先基于这些完整长度的歌曲在歌曲级别进行计算，然后对数据划分取均值，以获得最终测试值。

为评估滤波器泛化，我们使用已见滤波器和未见滤波器，对每个模型进行两项测试。如表 7.1所示，我们选择 6 阶巴特沃斯滤波器作为未见过滤器，因为它未用于任何训练设置。已见滤波器仅包括 6 阶切比雪夫一型，因为这是单滤波器和多滤波器训练设置中唯一共有的滤波器。

[2]https://github.com/tensorflow/models/tree/master/research/audioset/vggish

7.4.3 验证

为了观察整个训练过程中的泛化或过拟合情况，我们定期进行验证，每 2500 次训练迭代进行一次。我们使用 8 秒的音频片段进行验证，片段从每首歌曲的第 8 秒开始，且仅选取 8 首歌曲。这 8 首歌曲是 DSD100 训练集中的最后 8 首。由于训练过程中会反复进行验证，我们将验证集的样本量保持较小。我们认为较小的样本量足以满足需求，因为验证仅用于观察训练进程，最终的性能评估会在测试阶段进行。最终的信噪比值先通过计算每段 8 秒音频片段的信噪比得出，再对参与验证的歌曲取平均值。

与测试类似，验证也进行两次，分别使用已见和未见滤波器。使用未见滤波器进行验证时，采用与测试相同的 6 阶巴特沃斯滤波器。由于使用已见滤波器进行验证是为了观察各模型的训练进程，而非比较不同模型，因此所用滤波器与相应训练设置中的滤波器相同。如表 7.1所示，在单滤波器设置中，使用已见滤波器进行验证时仅使用切比雪夫 I 型滤波器；在多滤波器设置中，则使用全部八种训练滤波器，并将每种滤波器分别分配给验证数据划分中的一首不同歌曲进行处理。

7.5 实现细节

我们使用 PyTorch 框架（Paszke等人，2019年）和一块 Nvidia GeForce GTX 1080 Ti GPU 构建并训练模型。模型权重以随机方式初始化，其值取自均值为零、方差为一的正态分布。批量大小为 8。我们使用 Adam优化器（Kingma & Ba，2015年），初始学习率为 5e-4，beta 值为 0.9 和 0.999。当训练损失趋于平稳时，学习率减半。我们每 2500 次迭代记录一次平均训练损失，并将连续 5 次测量中损失均未下降视为趋于平稳。训练样本的生成方式是：先从训练数据集中随机选取一个音频文件，再在文件中的随机位置提取一段立体声音频数据块，长度为 8192 个样本，对应 186 毫秒。不过，由于所有模型均为全卷积模型，因此可以处理任意长度的音频信号。我们将模型训练至收敛；测试时使用训练结束时的模型权重。

7.6 结果

7.6.1 验证数据

图 7.3 概述了所有不同模型和训练方案的演奏表现，并展示了信噪比随训练迭代次数的变化情况。水平虚线表示输入信噪比水平的基线，其余曲线则表示两种验证设置下的输出信噪比。使用已见滤波器进行输入验证时的信噪比水平有所不同

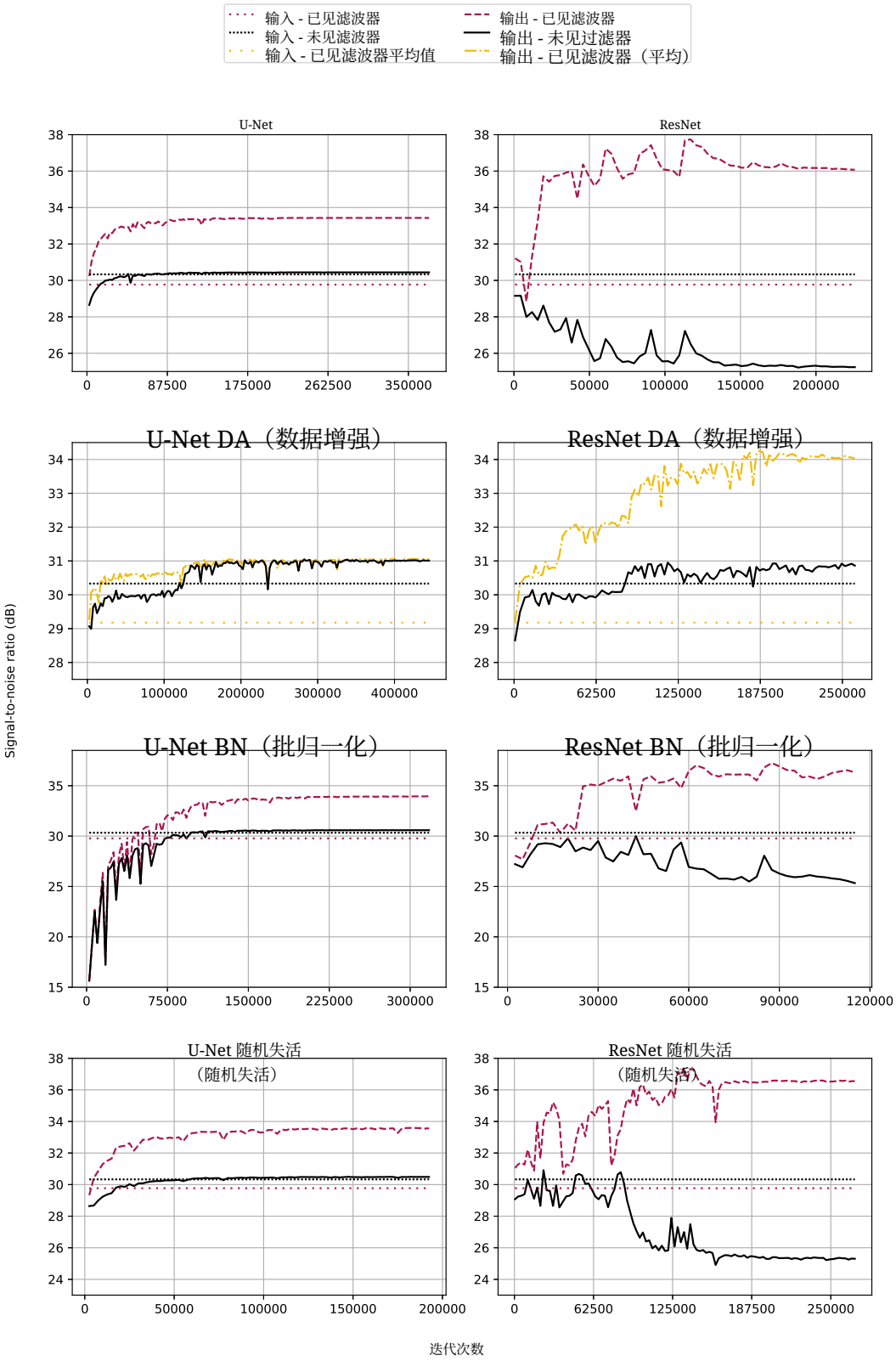


图 7.3：模型在训练过程中的验证性能。

由于使用了不同数量的训练滤波器，数据增强实验的情况如表 7.1 所总结，并如图 7.2 所示，这些滤波器不同的频率响应自然会导致不同的基线信噪比。

观察图 7.3 第一行可见，对于两个神经网络，当已知输入滤波器时，信噪比都可能显著优于基线。然而，对比 U-Net 和 ResNet 可以发现，它们在未见滤波器上的表现明显不同。对于 U-Net，输出信噪比会收敛至基线；而对于 ResNet，随着训练继续，演奏表现会下降。因此，我们可以清楚地看到，这两种模型都缺乏滤波器泛化能力。

接着看第二行，其中训练包含数据增强，我们可以观察到不同的模式：两个神经网络都能使未见滤波器的信噪比优于基线。对比 U-Net 和 ResNet 可以发现，ResNet 在已见滤波器集合上的提升幅度大于 U-Net，尽管两者的参数数量大致相同。

观察包含两种正则化技术的第三行和第四行，可以发现其模式与第一行大体相似：U-Net 再次收敛到输入信噪比，而 ResNet 得到的信噪比低于输入。总而言之，对于这两种神经网络，只有在训练中采用数据增强，我们才能获得高于输入的信噪比。

7.6.2 测试数据

如第 7.4.3 节所述，验证数据集规模较小，图 7.3 中展示的结果是对时值为 8 s 的短片段进行计算并取平均所得。在表 7.2 中，我们展示了模型在测试数据上的演奏表现，其中包括对整个测试数据集时值范围内的信噪比和作为感知度量的 VGG 距离的测量。 Δ信噪比和 $-\Delta$VGG 表示相对于输入的改进。对于信噪比， Δ信噪比和 $-\Delta$VGG 值越高越好，而 VGG 值越低越好。DA、BN 和 DO 分别表示数据增强、批归一化和随机失活。VGG 嵌入和 VGG 距离的取值范围为 0 到 255。

使用已见滤波器进行测试时，未采用数据增强的 ResNet 模型比所有 U-Net 变体至少高出 4 dB，较输入信噪比提升超过 7 dB。演奏表现最佳的模型是采用随机失活的 ResNet，较输入信噪比提升了 7.4 dB。我们还观察到，使用已见滤波器进行评估时，加入数据增强会降低演奏表现。

使用未见滤波器进行测试时，表现最好的两个模型都采用了我们提出的数据增强方法。在此情况下，未进行数据增强的 ResNet 变体所产生的输出信噪比远低于输入信噪比。其输出信噪比比输入低约 5.3 dB，比基于 U-Net 的对应模型低约 6.6 dB。加入数据增强后，基线 U-Net 和 ResNet 的演奏均有所提升。尽管 U-Net 的提升幅度较小，为 0.45 dB，但 ResNet 的提升幅度要大得多，达到 7.2 dB。使用未见滤波器进行测试时，表现最好的模型是采用数据增强的 ResNet，其输出信噪比比输入信噪比高 1.8 dB。

表 7.2：测试数据集上的输出信噪比（SNR）和绝对 VGG 距离（VGG），以及相对于输入的改进。

滤波器	实验信噪比		ΔSNR	VGG	−ΔVGG
切比雪夫1–6 (已见滤波器)	输入	27.86		46.55	
	U-Net	30.34	+2.47	41.04	+5.51
	U-Net DA	29.78	+1.91	44.29	+2.26
	U-Net BN	30.90	+3.03	41.52	+5.03
	U-Net 随机失活	30.49	+2.62	41.51	+5.04
	ResNet	34.94	+7.08	39.02	+7.53
	ResNet 数据增强	30.48	+2.62	40.11	+6.43
	ResNet 批归一化	34.37	+6.50	39.41	+7.14
	ResNet 随机失活	**35.27**	**+7.41**	**37.23**	**+9.32**
巴特沃斯–6 (未见过滤器)	输入	27.37		47.11	
	U-Net	28.55	+1.18	41.90	+5.21
	U-Net DA	29.00	+1.63	44.80	+2.31
	U-Net BN	28.77	+1.40	42.06	+5.06
	U-Net 随机失活	28.62	+1.24	42.34	+4.78
	ResNet	21.96	−5.41	47.12	−0.01
	ResNet 数据增强	**29.16**	**+1.78**	**40.52**	**+6.59**
	ResNet 批归一化	23.23	−4.14	46.38	+0.73
	ResNet 随机失活	22.10	−5.27	46.15	+0.96

从 VGG 距离来看，U-Net 变体的结果在不同滤波器之间变化不大。与使用已见滤波器的设置相比，未进行数据增强的 ResNet 变体在使用未见滤波器时结果较差；不过，这些值与输入值非常接近，因此就 VGG 距离而言，滤波器过拟合并不像信噪比那样严重。在使用未见滤波器的设置下，加入数据增强使 U-Net 的 VGG 距离恶化了 2.8，但使 ResNet 的 VGG 距离显著改善了 6.6，因此采用数据增强的 ResNet 在 VGG 距离和信噪比方面都是表现最好的模型。

每首测试歌曲的定量结果以及三个音频片段可在线查看[3]。

7.6.3 样本过拟合

表 7.3 分别展示了基线模型在训练和测试数据划分上的演奏表现；评估时使用了各数据划分中的全部样本，并覆盖其完整时值。输入由训练期间也使用的低通滤波器生成（已见滤波器，即 6 阶切比雪夫一型滤波器）。为了解是否发生了样本过拟合（即神经网络是否以某种方式记住了训练数据的音频内容），我们在测试期间使用已见滤波器，即 6 阶切比雪夫一型滤波器。对于基线 U-Net 和 ResNet，训练与测试数据划分之间相较于输入的信噪比提升表明，模型并未对音频样本本身发生过拟合。

[3] https://serkansulun.com/bwe

表7.3：未采用任何正则化方法时，我们的基线模型在训练和测试数据划分上的输出信噪比（SNR），以及与输入信噪比的比较。

Data 划分	实验信噪比（dB）		改进（dB）
训练	输入	25.99	
	U-Net	28.34	+2.35
	ResNet	33.00	+7.01
测试	输入	27.86	
	U-Net	30.34	+2.48
	ResNet	34.94	+7.08

7.6.4 U-Net 与 ResNet：模型比较

按照第 7.1 节中的规定，我们让 U-Net 和 ResNet 的参数数量大致相当。不过，我们非正式地观察到训练时间存在明显差异。表 7.4 展示了这些神经网络的若干客观属性，即基线模型的参数数量、乘加运算次数（MACs）和运行时间。MACs 数量大致相当于浮点运算次数（FLOP）的一半。运行速率指测试期间处理长度为一秒的信号所需的时间（以秒计），也就是不计算梯度的前向传播。尽管两个模型的参数数量大致相同，但 U-Net 的运行时间要短得多，乘加运算次数也更少。这是由于其类似自编码器的结构：通道较多的卷积层位于神经网络的瓶颈附近，此处的空间激活尺寸最小，因此有效减少了乘加运算次数和运行时间。再次观察图 7.3，我们可以推测，ResNet 的表征能力强于 U-Net，因为它比 U-Net 更能对多个已知滤波器进行建模，尽管代价是训练和推理速度较慢。

表 7.4：我们模型的参数数量、乘加运算次数（MACs）和运行时间。

模型	参数数量	数量 MACs	运行时间 rate
U-Net	56.4M	415.3G	0.14
ResNet	55.1M	3609.4G	1.06

7.6.5 带宽扩展的可视化

尽管我们提出的方法完全在时域中运行，但我们仍以图形方式概述两个神经网络的输出，对比基线版本与加入数据增强后的版本，并分别考察已见滤波器和未见滤波器。为此，我们在图 7.4 中展示了测试集在上述各种条件下的一个音频片段的声谱图。观察该图可发现，与 ResNet 相比，U-Net 的行为截然不同。总体而言，我们可以看到 ResNet 的输出中有更显著的高频信息。尤其值得注意的是

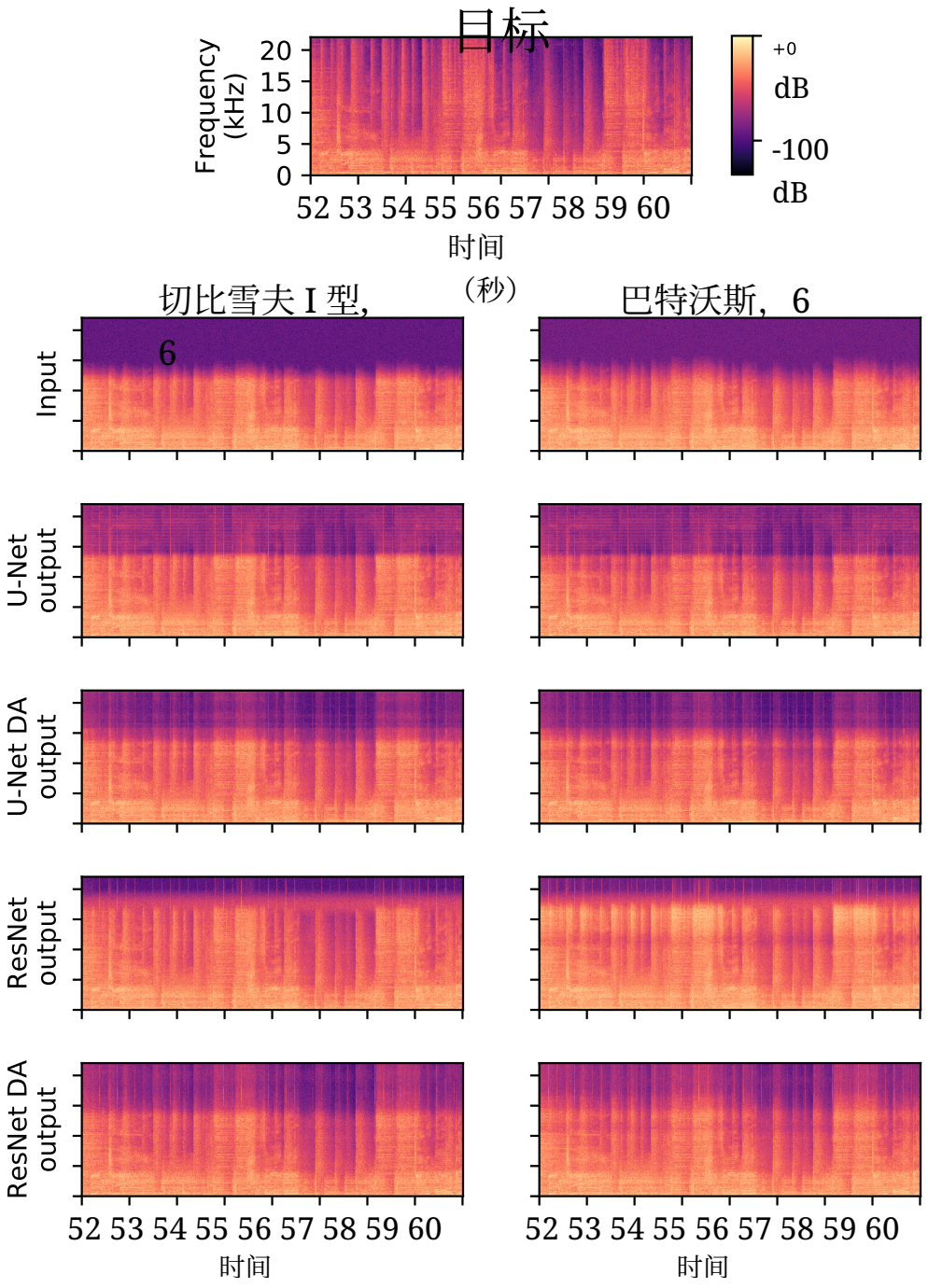


图7.4：样本音频片段的频谱图。

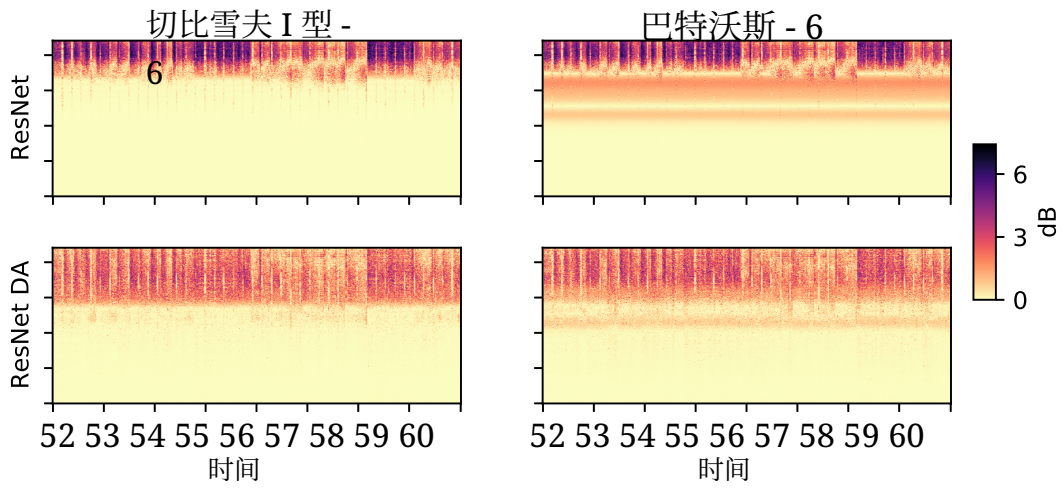


图7.5：与目标频谱图的绝对差值。为提高可见度，反转了色彩图。

基线残差网络和未见过的巴特沃斯滤波器在约 12–17 kHz 的频率区域中存在差异；与目标相比，后者似乎对该区域进行了“过度增强”。相比之下，加入数据增强后，这种高频提升便不再明显。为进一步突出这一现象，我们在图7.5 中展示了基线残差网络以及使用数据增强的残差网络与目标频谱图之间的绝对差值。对于未见过的巴特沃斯滤波器，在频谱图的上半部分，使用数据增强的残差网络的绝对差值比基线残差网络平滑得多。从这一视觉表征中可以清楚地看到，在所有条件下，绝对差值频谱图的下半部分基本保持不变，这反映了网络架构中输入与输出之间直接的加性连接。

7.7 讨论

我们提出了一个问题：应用于音乐音频带宽扩展的深度神经网络能否实现滤波器泛化。与深度学习所应用的许多问题不同，我们没有发现任何证据表明模型会对音频样本本身（即训练数据）过拟合；相反，我们观察到一个明显趋势：当前最先进的深度神经网络会对滤波器形状过拟合。当这些深度神经网络处理经训练时未使用的低通滤波器预处理的音频片段时，无法实现有意义的带宽扩展。此外，采用批归一化和随机失活等广泛使用的正则化层，也不足以缓解这一问题。着眼于更广泛的上下文以及音频增强中音乐音频带宽扩展的长期目标，我们认为滤波器过拟合是一个关键问题，值得持续关注。

为了解决滤波器过拟合问题，我们提出了一种新颖的数据增强方法，在训练时使用多个滤波器。结果表明，如果不进行数据增强，滤波器过拟合会随着训练的推进而加剧；而采用数据增强，则是实现滤波器泛化的一个有希望的步骤。虽然 U-Net 的泛化能力提升幅度较小，但 ResNet 的效果更为显著，它在多个已见训练滤波器上都保持了较高的演奏表现。尤其值得注意的是，不采用数据增强的 ResNet 变体在使用未见滤波器进行测试时，结果很差，输出质量远低于输入。因此，采用数据增强是唯一能够使信噪比高于输入水平的方法。

除了关于滤波器泛化的主要发现外，据我们所知，这是音频处理领域首次对 U-Net 和 ResNet 进行比较；或许也是首次在参数数量相近的情况下比较这两种方法。考察使用已见滤波器进行测试的结果，我们发现基线残差网络的表现远胜基线 U-Net。然而，使用未见过滤器进行测试时，基线残差网络的表现最差。

我们认为，ResNet 的表征能力强于 U-Net，因为 U-Net 会在其下采样模块中缩小空间激活尺寸，而 ResNet 则从输入到输出始终保持空间激活尺寸不变，从而最大限度地减少信息损失。尽管这两个神经网络的参数数量相同，但我们可以通过比较乘累加运算次数来量化这种更强的表征能力。表征能力更强，使 ResNet 在使用已见滤波器进行测试时表现出色；但在不使用数据增强时，也表现出更严重的滤波器过拟合。我们证明，采用所提出的数据增强方法，可以成功地对这个强大的神经网络进行正则化，并使其在使用未见过滤器进行测试时获得最高的信噪比。虽然我们选择让两个模型的参数数量大致相等，但值得注意的是，与 ResNet 相比，U-Net 的速度快.5 倍，乘累加运算（MACs）次数则少近 9 倍。因此，U-Net 可能是实时流应用的更优架构。

现在我们进入结论部分，总结本研究的主要发现及其意义，并展望未来可能的研究方向。

第8章

结论

在本论文中，我们提出了一种新颖的基于视频的音乐生成方法，并打造了 EMSYNC——一个创新框架，通过对齐情绪内容和时间边界，使音乐与视频同步。EMSYNC 将视频分析、基于情绪的条件控制和时间边界条件控制整合到单一的音乐生成流程中，为生成在情绪和节奏上与视频相匹配的音乐这一关键挑战提供了解决方案。

8.1 贡献

本研究的一项主要贡献是创建了用于任意视频情绪分类以及电影预告片流派分类的模型。这些模型是我们的基于视频的音乐生成系统取得成功的关键，因为它们提供了音乐生成所需的情感语境和类型特征。我们的情感分类器 VEMOCLAP 能够识别任意视频的情感基调，因此除了音乐生成之外，还可作为适用于各种应用的多功能工具。同样，我们针对电影预告片的流派分类器利用预训练模型进行特征提取，从而实现高效、高质量的流派分类。这些预训练模型是对社区颇有价值的资源，在多媒体内容分析领域具有广泛的应用潜力，包括自动内容分类、情感分析和推荐系统。

另一项重要贡献是我们的音乐生成器，它首次采用连续值条件——具体而言，效价和唤醒度值——而非离散情感类别进行条件控制。这一突破使我们能够对生成音乐进行更精细的控制，因为它可以精准调节情感表达和音乐动态。通过避免对情感特征进行离散化，我们保留了情感表达的连续性，从而生成更细腻、能更充分体现视频内容中情感复杂性的音乐。这种连续条件控制方法为微调音乐生成提供了更大的灵活性，能够生成更具表现力、适应性更强的输出，更好地契合不同视频中变化的情感基调。此外，这种方法也对其他

序列生成模型具有启发意义，尤其是在自然语言处理等领域，对生成内容进行细粒度控制有助于提升生成序列的连贯性和情感深度。

我们还创建了首个专为音乐生成设计的大规模情感标注 MIDI 数据集。该数据集结合了从 Spotify 开发者 API 和歌曲歌词中提取的情感标注，为训练基于情感的音乐生成器提供了强大的资源。这个情感标注数据集影响深远，它通过提供丰富的情感训练数据，弥补了符号音乐生成领域的一项空白，使生成的音乐能够明确地契合不同的情绪状态。该数据集为情感音乐生成研究开辟了新机遇，使人们能够创建可根据不同情境中的情感输入动态作出响应的系统，例如电影、游戏和交互式媒体。

此外，作为这项工作的一部分创建的带情感标签的 MIDI 数据集，对人工智能和音乐领域具有更广泛的意义。该数据集是训练系统的独特资源，这些系统不仅能够生成音乐，还能理解并适应一段乐曲的情感特质。通过同时提供连续情感标签和符号音乐数据，我们为探索更先进的音乐生成系统奠定了基础，使其能够以自然、流畅的方式响应情感线索。

我们的工作还通过引入一种利用时间边界条件控制来同步视频内容与音乐的系统，作出了重要贡献。通过将音乐事件与视频场景切换对齐，我们的方法确保音乐不仅在情感上与视频相契合，也能在节奏上与视频的关键时刻同步。这种同步提升了整体多媒体体验，尤其适用于电影配乐、广告和电子游戏等精确时序至关重要的场景。

虽然 MIDI 事件本身并未与固定时间网格对齐，我们的方法仍然能够对生成器进行有效的时间条件控制。这对于涉及 Transformer 的序列处理任务具有更广泛的意义，因为它展示了一种将时间边界融入生成模型的新方法。这种方法还可扩展到文本生成、语音合成及其他基于序列的任务；在这些任务中，精确的时间对齐对于维持连贯性和提升整体输出质量至关重要。

总之，EMSYNC 代表着视频音乐生成领域的一项重大进展。它是首个将基于情感和基于边界的条件控制整合到音乐生成中的系统，能够创作出情感丰富且在节奏上与视频同步的音乐。预训练的情感和流派分类器、音乐生成中的连续条件控制方法，以及情感标注 MIDI 数据集，都是重要创新，为开发更先进、更多样化的音乐生成系统铺平了道路。

此外，在用户研究中，无论参与者自我报告是否具备乐理知识，EMSYNC 在所有主观指标上均优于现有的两种方法。来自不同音乐熟悉程度的参与者一致认为，我们的方法在音乐丰富度、质量、情感对齐和时序同步方面表现最佳

。这不仅展示了 EMSYNC 的技术效率，也体现了它重塑基于视频的音乐生成的潜力，为音乐与视频内容的同步提供了更具表现力和适应性的解决方案。这些研究结果进一步验证了 EMSYNC 的实际效果，使其成为该领域一项值得瞩目的进展。

最后，我们通过音频带宽扩展任务对音频生成进行了初步分析。我们指出了使用合成数据所带来的一个重要问题，这种做法在许多音频生成领域的研究中都很常见。具体而言，我们证明，使用一种数据合成方法训练的模型，无法很好地泛化到由其他合成方法生成的数据。为了解决这一问题，我们率先尝试通过多样化数据合成方法，更好地模拟真实世界场景，以缓解这一问题。我们相信，本研究的见解对其他音频生成任务也具有价值，包括 MIDI 转音频。

我们通过开源代码、数据集、预训练模型和交互式演示，让社群能够接触到研究的每个阶段。发布这些资源，旨在促进合作与创新，受益者不仅包括研究社群，也包括广大公众。这种开放性使其他研究人员能够以我们的工作为基础，测试新想法，并推动基于视频的音乐生成领域的发展。此外，开放这些工具也让从业者、开发者和爱好者能够在各种真实场景中探索并应用我们的方法，应用范围涵盖电影制作、游戏开发等领域。我们希望通过这项贡献，激发学术界和创意领域进一步探索并取得进展。

本论文的主要贡献总结如下。

- 我们为 LakhMIDI 数据集中的 34,791 首歌曲提供了效价-唤醒度值形式的情绪标签（Raffel，2016），由此创建了现有规模最大的带有情绪标签的 MIDI 数据集。

- 我们提出了首个能够以连续效价-唤醒度值为条件的 MIDI 生成器。

- 我们创建了首个能够以直接方式将时间边界作为条件的 MIDI 生成器。

- 我们的电影预告片类型分类器在 MovieNet（规模最大的电影预告片数据集）上的表现优于所有现有方法（Q. Huang 等人，2020）。

- 我们的视频情绪分类器在 Ekman-6（规模最大的带有情绪标签的用户生成（任意）视频数据集）上的表现优于所有现有方法（B. Xu 等人，2018年）。

- 我们的基于视频的音乐生成器在用户研究中，各项指标和各类人群的表现均优于所有现有方法。

- 我们证明，使用合成数据训练音频生成模型会导致泛化问题，并初步着手应对这一挑战。

下面列出我们在研究期间发表的论文。

• **S. Sulun**、M. E. P. Davies 和 P. Viana，《基于连续值情绪条件的符号音乐生成》，IEEE Access，第 10 卷，第 44617–44626 页，2022年。

• **S. Sulun**、P. Oliveira 和 P. Viana，《Emotion4MIDI：基于歌词且带情绪标签的符号音乐数据集》，载于《Progress in Artificial Intelligence》，Cham：Springer Nature Switzerland，2023年，第77–89页。

• **S. Sulun**、P. Viana 和 M. E. P. Davies，《通过对齐情感与时间边界生成视频配乐》，《IEEE多媒体汇刊》，2026。 [印刷中]。

• **S. Sulun**、P. Viana 和 M. E. P. Davies，《VEMOCLAP——一个视频情感分类Web应用》，IEEE国际多媒体研讨会（ISM），2024。

• **S. Sulun**、P. Viana 和 M. E. P. Davies，《使用多模态预训练特征进行电影预告片体裁分类》，《专家系统与应用》，第258卷，第125209页，2024。

• **S. Sulun** 和 M. E. P. Davies，“基于深度神经网络的音乐带宽扩展中滤波器的泛化”，IEEE 信号处理选题期刊，第 15 卷，第 1 期，第 132–142 页，2020年。

8.2 未来方向

展望未来，我们希望为机器学习和多媒体领域的后续研究者提供见解与启发。在整个研究过程中，我们倾向于采用大型且相对简单的模型，并使用大型数据集进行训练，而不是采用复杂的小型数据集训练模型。我们方法在定性和定量方面的优势，与当前趋势相符：在更大型的数据集上训练更大型的模型，能够取得更好的结果（Bernstein 等人，2021；Sevilla 等人，2022年；M. Tan 和 Le，2019年）。行业长期以来一直采用这种方法，文本生成模型包括 ChatGPT[1], LLaMA[2], 和 Gemini[3], ，图像生成模型包括 DALL-E[4] 和 Imagen[5], ，音频生成模型则包括 Suno[6]。

尽管大型科技公司的硬件能力让它们拥有明显优势，但另一个常被忽视的因素是它们能够获取大规模专有数据集。因此，我们强调创建更新、更大规模的开源数据集和经同行评审的数据集的重要性，而不是一味地专注于在小型数据集上构建复杂模型，以在现有最先进水平的基础上取得渐进式改进。

创建更优质数据集的一种有效方法，是为现有数据集添加标签。CLIP 等图像处理模型之所以成功，源于对比式预训练方法：该方法将图像与从互联网抓取的标签及其他文本描述配对（Radford 等人，2021）。

[1] https://www.chatgpt.com [2] https://www.llama.com [3] https://gemini.google.com/app [4] https://openai.com/index/dall-e-3 [5] https://deepmind.google/technologies/imagen-3 [6] https://suno.com

类似的方法也可以应用于其他类型的多媒体。不过，这种方法面临的一项关键挑战是有意或无意地使用受版权保护的材料。为了跟上行业研究进展，学术实验室似乎很可能需要与法律团队合作，或拨出研究经费用于法律咨询，就像行业所做的那样。

许多数据集依赖视频分享平台上的内容（Q. Huang 等人，2020年；B. Xu 等人，2018年）。大语言模型的兴起使文本数据尤为重要。新一代数据集应充分利用用户上传视频的标签和描述。对比学习策略在从视频和文本配对中提取有意义的特征方面尤其有用（Alayrac 等人，2020年）。

当所分析的视频具有复杂的情节、语义变化和微妙之处时，视频情感分类会更具挑战性。直接从原始像素和音频中提取这些复杂的语义信息，效率可能很低，甚至难以实现。我们采用预训练特征的方法，是朝着高效理解这种语义复杂性迈出的一步。下一步，可以利用类似的预训练特征将视频详细转录为文本，并由具备推理能力的大语言模型（LLM）（Wei 等人，2022年）处理生成的文本，以检测潜在的情节转折和细微之处。

数据稀缺时，使用高维表示可能会导致过拟合，这种现象通常称为维度灾难。我们使用预训练特征的方法实际上是一种有效的特征降维技术，能够将大量原始帧压缩为数量较少且更有意义的特征。当然，也可以添加其他特征。例如，我们原计划纳入语音情绪分类这一方面，但难以找到性能足够好的开源模型。不过，尽管预训练特征提取器是在推理模式下运行的，使用多个提取器仍会显著增加运行时间。为缓解这一问题，可以采用知识蒸馏来缩小这些预训练网络的规模（Hinton等人，2015年）。此外，将多种架构整合为一个统一模型，将是一项颇具挑战但也很有价值的研究任务。最后，通过持续学习利用来自不同分类的多个数据集，或许有助于解决数据稀缺问题（L. Wang等人，2024）。

带标签的 MIDI 数据集同样存在数据稀缺问题。我们的工作利用 Spotify Developer API 中的特征为 MIDI 文件添加标签。由于这些特征最初是为音频而非 MIDI 创建的，因此我们将其称为“弱标签”。不过，我们的研究表明，即使标签较弱，拥有包含这些标签的大型数据集仍然有益。未来的研究者可以从互联网上收集与音乐音频相关的文本特征，例如描述或标签，将这些特征关联到相应的 MIDI 文件，并以类似方式应用大语言模型和/或对比学习方法。

使用深度神经网络（DNN）从 MIDI 生成多乐器音频，或反向将音频转换为 MIDI，是一个相对少有人探索的领域。高效的音频到 MIDI 模型有助于合成生成大型 MIDI 数据集。另一方面，输出逼真的 MIDI 转音频模型则能以符号音乐作品为基础，制作出高质量音频。扩散模型已展现出生成前所未有的图像的能力，

质量，因而也可能成为生成高质量、逼真音频的理想选择（Sohl- Dickstein 等，2015年）。

最后，Suno 的出现无疑代表着音乐生成领域的一项颠覆性突破。尽管其方法属于专有技术，但可以合理推测，大规模的私有音频格式音乐数据集为其成功作出了重要贡献。若能提出一种开源、经过同行评审且取得类似成功的 方法，将标志着音乐生成研究取得了非凡进展。

参考文献

Adlam, B., Weill, C., & Kapoor, A.（2019年）。探究 Wasserstein生成对抗网络中的欠拟合与过拟合。CoRR，abs/1910.14137。http://arxiv.org/abs/1910.14137 Adler, A., Emiya, V., Jafari, M. G., Elad, M., Gribonval, R., & Plumbley, M. D.（2012年）。音频修复。I EEE语音与音频处理汇刊，20(3)，922–932。https://doi.o rg/10.1109/TASL.2011.2168211 Ak, K. E., Lee, G.-G., Xu, Y., & Shen, M.（2023年）。在 Transformer 中运用高效训练与特征融合进行多模态分类。IEEE国际图像处理会议，ICIP 2023，马来西亚吉隆坡，2023年10月8日至11日，1420–1424。https: //doi.org/10.1109/ICIP49359.2023.10223098 Alayrac, J., Recasens, A., Schneider, R., Arandjelovic, R., Ramapuram, J., Fauw, J. D., Smaira, L., Dieleman, S., & Zisserman, A.（2020年）。自监督多模态通用网络。神经信息处理系统进展。https://proceedings.neurips.cc/pape r/2020/hash/0060ef47b12160b9198302ebdb144dcf-Abstract.html Almahairi, A., Ballas, N., Cooijmans, T., Zheng, Y., Larochelle, H., & Courville, A. C.（2016）。动态容量网络。第33届国际机器学习会议论文集，ICML 2016，美国纽约市，纽约州，2016年6月19日至24日，48，2549–2558。http://proceedings.mlr.press/v48/almahairi16.html Almeida, A., de Villiers, J. P., De Freitas, A., & Velayudan, M.（2022年）。从视频内容中提取的多种深度学习特征对视频推荐的互补性。《专家系统与应用》，192，116335。https://doi.org/10.1016/j.eswa.20 21.116335 Almeida, J., Vilaça, L., Teixeira, I. N., & Viana, P.（2021）。通过面部表情识别电影中的情绪。《应用科学》，11(15)。https : / / doi . org / 10 . 33 90/app11156827 Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C.（2021）。ViViT：视频视觉Transformer。2021 IEEE/CVF 国际计算机视觉会议，ICCV 2021，加拿大魁北克省蒙特利尔，2021年10月10日至17日，6816–6826。https://doi.org/10.1109 /ICCV48922.2021.00676 Ba, L. J., Kiros, J. R., & Hinton, G. E.（2016）。层归一化。CoRR，abs/1607.06450。http: //arxiv.org/abs/1607.06450 Bahdanau, D., Cho, K., & Bengio, Y.（2015）。通过联合学习对齐与翻译实现神经机器翻译。第三届国际学习表征会议，ICLR 2015，美国加利福尼亚州圣迭戈，2015年5月7日至9日，会议论文集。http://arxi v.org/abs/1409.0473 Bain, M., Nagrani, A., Brown, A., & Zisserman, A.（2020年）。浓缩电影：基于故事、利用上下文嵌入进行检索。计算机视觉——ACCV 2020——第15届亚洲计算机视觉会议，日本京都，2020年11月30日至12月4日，修订版

精选论文，第五部分，12626，460–479。https://doi.org/10.1007/978-3-030-69541-5_28 Bansal, D., Raj, B., & Smaragdis, P.（2005）。利用非负矩阵分解扩展窄带语音的带宽。第九届欧洲语音通信与技术会议，INTERSPEECH-Eurospeech 2005，葡萄牙里斯本，2005年9月4日至8日，1505–1508。https://doi.org/10.21437/INTERSPEECH.2005-528 Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z.（2016）。利用众包标签分布训练深度神经网络进行面部表情识别。第18届ACM多模态交互国际会议论文集，279–283。https://doi.org/10.1145/2993148.2993165 Bauer, P., & Fingscheidt, T.（2008）。一种通过跨语言训练与测试评估的基于隐马尔可夫模型的人工带宽扩展。IEEE声学、语音与信号处理国际会议论文集，ICASSP 2008，2008年3月30日至4月4日，美国内华达州拉斯维加斯凯撒宫，4589–4592。https://doi.org/10.1109/ICASSP.2008 .4518678 Berg, J., & Wingstedt, J.（2005）。特定音乐参数与表达情绪之间的关系：拓展计算机娱乐的潜力。国际计算机娱乐技术进展会议论文集，ACE 2005，西班牙瓦伦西亚，2005年6月15日至15日，164–171。https://doi.org/10.1145/1178477.1178499 Bernstein, L., Sludds, A., Hamerly, R., Sze, V., Emer, J., & Englund, D.（2021）。适用于深度学习的可自由扩展、可重构光学硬件。《科学报告》，11(1)，3144。htt ps://doi.org/10.1038/s41598-021-82543-3 Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P.（2011）。百万歌曲数据集。第12届国际音乐信息检索学会会议论文集，ISMIR 2011，美国佛罗里达州迈阿密，2011年10月24日至28日，591–596。http://ismir2011.ismi r.net/papers/OS6-1.pdf Bharucha, J., & Krumhansl, C. L.（1983）。音乐中和声结构的表征：稳定性层级随上下文变化。《认知》，13(1)，63–102。https://doi.org /10.1016/0010-0277(83)90003-3 Bieda, I., Kisil, A., & Panchenko, T.（2021）。一种场景变化检测方法。2021年第11届IEEE智能数据采集与高级计算系统：技术与应用国际会议（IDAACS），波兰克拉科夫，2021年9月22日至25日，489–493。https://doi.org/10.1109/IDAACS53288.2021.9660887 Bittner, R. M., Wilkins, J., Yip, H., & Bello, J. P.（2016）。Medleydb 2.0：用于可持续数据收集的新数据与系统。ISMIR最新研究与演示论文。https://bpb-us-e1 .wpmucdn.com/wp.nyu.edu/dist/2/2294/files/2016/08/bittner-medleydb.pdf?bid=2294Bochkovskiy, A., Wang, C.-Y., & Liao, H.-Y. M.（2020年4月）。YOLOv4：目标检测的最佳速度与准确率。https://doi.org/10.48550/arXiv.2004.10934 Böck, S., & Widmer, G.（2013）。用于起音检测的最大滤波器颤音抑制。第16届国际数字音频效果会议（DAFx），7，4。Bose, D., Hebbar, R., Somandepalli, K., Zhang, H., Cui, Y., Cole-McLaughlin, K., Wang, H., & Narayanan, S.（2023）。MovieCLIP：电影中的视觉场景识别。IEEE/CVF计算机视觉应用冬季会议，WACV 2023，美国夏威夷州威可洛亚，2023年1月2日至7日，2082–2091。https://doi.org/10.1109/WACV56688.2023.00212 Briot, J., Hadjeres, G., & Pachet, F.（2020）。用于音乐生成的深度学习技术。施普林格。https://doi.org/10.1007/978-3-319-70163-9 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child,

R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D.（2020）。语言模型是少样本学习器。《神经信息处理系统进展》33：神经信息处理系统年会2020，NeurIPS 2020，2020年12月6日至12日，线上举行。https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb49674 18bfb8ac142f64a-Abstract.html Bucila, C., Caruana, R., & Niculescu-Mizil, A.（2006）。模型压缩。第12届ACM SIGKDD知识发现与数据挖掘国际会议论文集，美国宾夕法尼亚州费城，2006年8月20日至23日，535–541。https://doi.org/10.1145/11504 02.1150464 Buechel, S., & Hahn, U.（2017）。EmoBank：研究标注视角和表征格式对维度情感分析的影响。计算语言学协会欧洲分会第15届会议论文集，EACL 2017，西班牙瓦伦西亚，2017年4月3日至7日，第2卷：短论文，578–585。https://doi.org/10.18 653/V1/E17-2092 Buhler, J.（2018）。电影配乐理论。牛津大学出版社。https://books.google.com /books?id=n_NwDwAAQBAJ&lpg=PP1&ots= RKwzh_dMet&lr=&pg=PP1&redir_esc=y#v=onepage&q&f=false Calvo, R. A., & Mac Kim, S.（2013）。文本中的情感：维度模型与类别模型。计算智能，29(3)，527–543。https://doi.org/10.1111/j.1467-8640.2012.0 0456.x Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Anke, L. E., Liu, F., & Cámara, E. M.（2022）。TweetNLP：面向社交媒体的前沿自然语言处理。2022年自然语言处理实证方法会议论文集，EMNLP 2022——系统演示，阿布扎比，阿联酋，2022年12月7日至11日，38–49。https://doi.org/10.18653/V1/2022.EMNLP-DEMOS .5 Canny, J. F.（1986）。边缘检测的计算方法。IEEE模式分析与机器智能汇刊，8(6)，679–698。https://doi.org/10.1109/TPAMI.1986.476 7851 Carreira, J., & Zisserman, A.（2017）。动作识别将走向何方？一种新模型与Kinetics数据集。2017年IEEE计算机视觉与模式识别会议，CVPR 2017，美国夏威夷州檀香山，2017年7月21日至26日，4724–4733。https://doi.org/10.1109/CVPR.2017.5 02 Cascante-Bonilla, P., Sitaraman, K., Luo, M., & Ordonez, V.（2019）。MovieScope：利用多模态对电影进行大规模分析。arXiv预印本 arXiv:1908.03180。https://d oi.org/10.48550/arXiv.1908.03180 Chatterjee, A., Narahari, K. N., Joshi, M., & Agrawal, P.（2019）。SemEval-2019任务3：EmoContext文本上下文情感检测。第13届语义评测国际研讨会论文集，SemEval@NAACL-HLT 2019，美国明尼苏达州明尼阿波利斯，2019年6月6日至7日，39–48。https://doi.org/10.18653/V1/S19-2005 Chen, C., Wu, Z., & Jiang, Y.-G.（2016）。上下文中的情感：用于视频情感识别的深层语义特征融合。ACM-MM 2016，127–131。https://doi.org/10.1145/2964284.2 967196 Chen, K., Wang, C., Berg-Kirkpatrick, T., & Dubnov, S.（2020）。Music SketchNet：通过音高与节奏的因子化表征实现可控音乐生成。第21届国际音乐信息检索学会会议论文集，ISMIR 2020，加拿大蒙特利尔，2020年10月11日至16日，77–84。http : / /archives . ismir . net / ismir2020 / paper/000146.pdf

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., & Wei, F. (2023). BEATs：使用声学分词器进行音频预训练。国际机器学习会议，ICML 2023，2023年7月23日至29日，美国夏威夷州檀香山，202，5178–5193。https: // proceedings.mlr.press/v202/chen23ag.html Cheng, Y. M., O'Shaughnessy, D. D., & Mermelstein, P. (1994). 从窄带语音中统计恢复宽带语音。IEEE语音与音频处理汇刊，2(4)，544–548。https://doi.org/10.1109/89.326637 Chennoukh, S., Gerrits, A. J., Miet, G.,& Sluijter, R. J. (2001). 利用线谱频率进行频带扩展以增强语音。IEEE声学、语音与信号处理国际会议，ICASSP 2001，2001年5月7日至11日，美国犹他州盐湖城盐宫会议中心，论文集，665–668。https://doi.org/10 .1109/ICASSP.2001.940919 Chiou, E., Giganti, F., Bonet-Carne, E., Punwani, S., Kokkinos, I., & Panagiotaki, E. (2018). 使用卷积神经网络对VERDICT DW-MRI中的前列腺癌进行分类。医学影像中的机器学习——第九届国际研讨会，MLMI 2018，与MICCAI 2018同期举行，西班牙格拉纳达，2018年9月16日，论文集，11046，319–327。https://doi.org/10.1007/978-3-030-00919-9_37 Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014a). 使用循环神经网络编码器—解码器学习用于统计机器翻译的短语表征。2014年自然语言处理实证方法会议论文集，EMNLP 2014，2014年10月25日至29日，卡塔尔多哈，SIGDAT会议，ACL的一个特别兴趣小组，1724–1734。https://doi.org/10.3115/v1/d1 4-1179 Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014b). 关于神经机器翻译的特性：编码器—解码器方法。SSST@EMNLP 2014论文集，第八届统计翻译中的句法、语义与结构研讨会，卡塔尔多哈，2014年10月25日，103–111。https://doi.org/10.3115/V1/W14-4012 Choi, K., Fazekas, G., & Sandler, M. B. (2016). 使用深度卷积神经网络进行自动标注。第17届国际音乐信息检索学会会议论文集，ISMIR 2016，美国纽约市，2016年8月7日至11日，805–811。https://wp.nyu.edu/ismir2016/wp-content/uploads/sites/2294/2016/07/009%5C_ Pap er.pdf Choi, K., Hawthorne, C., Simon, I., Dinculescu, M., & Engel, J. H. (2020). 使用Transformer自编码器编码音乐风格。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，线上会议，119，1899–1908。http://p roceedings.mlr.press/v119/choi20b.html Church, K. W., Chen, Z., & Ma, Y. (2021). 新兴趋势：微调简介。《自然语言工程》，27(6)，763–778。https://doi.org/10.1017/S1351324921000322 Cohen, A. J. (1990). 理解音乐配乐。《艺术实证研究》，8(2)，111–124。https://doi.org/10.2190/8Y6G-KTM8-VDX4-UHRW Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). 大规模无监督跨语言表征学习。第58届计算语言学协会年会论文集，ACL 2020，在线，2020年7月5日至10日，8440–8451。https://doi.org/10.1 8653/V1/2020.ACL-MAIN.747 Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. (2018). XNLI：评估跨语言句子表示。2018年

自然语言处理实证方法会议，布鲁塞尔，比利时，2018年10月31日至11月4日，2475–2485。https://doi.org/10.18653/V1/D18-1269 Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., & Défossez, A. (2023). 简单且可控的音乐生成。神经信息处理系统进展36：2023年神经信息处理系统年会，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日。http://papers.nips.cc/paper%5C_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., Sun, A. Y., Wang, S., Wenzek, G., Youngblood, A., Akula, B., Barrault, L., Gonzalez, G. M., Hansanti, P., Hoffman, J., ... Wang, J. (2022). 不让任何语言掉队：扩展以人为中心的机器翻译。https://doi.org/10.48550/ARXIV.2207.04672 Cutler, R., Saabas, A., Pärnamaa, T., Purin, M., Indenbom, E., Ristea, N., Guzvin, J., Gamper, H., Braun, S., & Aichner, R. (2023). ICASSP 2023声学回声消除挑战。CoRR，abs/2309.12553。https://doi.org/10.48550/ARXIV.2309.12553 Cutting, J. E. (2005). 感知电影与现实世界中的场景。《动态影像理论：生态学考量》，9–27。https://www.researchgate.net/profile/James-Cutting/publication/237009484_Perceiving_scenes_in_film_and_in_the_world/links/0046351ae10f2baa8c000000/Perceiving-scenes-in-film-and-in-the-world.pdf Dathathri, S., Madotto, A., Lan, J., Hung, J., Frank, E., Molino, P., Yosinski, J., & Liu, R. (2020). 即插即用语言模型：一种可控文本生成的简单方法。第八届国际表征学习会议，ICLR 2020，埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日。https://openreview.net/forum?id=H1edEyBKDS Defernez, M., & Kemsley, E. K. (1999). 利用人工神经网络（ANNs）避免高维数据分析中的过拟合。《分析家》，124(11)，1675–1681。https://doi.org /10.1039/A905556H Deldjoo, Y., Constantin, M. G., Ionescu, B., Schedl, M., & Cremonesi, P. (2018). MMTF-14K：用于推荐与检索的多维度电影预告片特征数据集。第九届ACM多媒体系统会议论文集，MMSys 2018，荷兰阿姆斯特丹，2018年6月12日至15日，450–455。https://doi.org/10.1145/3204949.3208141 Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A. S., Nemade, G., & Ravi, S. (2020). GoEmotions：一个细粒度情绪数据集。第58届计算语言学协会年会论文集，ACL 2020，在线，2020年7月5日至10日，4040–4054。https://doi.org/10.18653/V1/2020.ACL-MAIN.372 Deng, J., Dong, W., Socher, R., Li, L., Li, K., & Fei-Fei, L. (2009). ImageNet：大规模分层图像数据库。2009年IEEE计算机学会计算机视觉与模式识别会议（CVPR 2009），2009年6月20日至25日，美国佛罗里达州迈阿密，248–255。https://doi.org/10.1109/CVPR.2009.5206848 Deutsch, S. (2007). 配乐（让音乐各得其所）。《配乐》，1(1)，3–13。http: //eprints.bournemouth.ac.uk/1307/ Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT：用于语言理解的深度双向Transformer预训练。2019年计算语言学协会北美分会人类语言技术会议论文集，NAACL-HLT 2019，美国明尼苏达州明尼阿波利斯，2019年6月2日至7日，第1卷（长短篇论文），4171–4186。https://doi.org/10.18653/v1/n19-1423

Di, S., Jiang, Z., Liu, S., Wang, Z., Zhu, L., He, Z., Liu, H., & Yan, S. (2021). 视频背景音乐生成与可控音乐 Transformer. MM '21：ACM多媒体会议，线上会议，中国，2021年10月20日至24日，2037–2045。https://doi.org/10.1145 /3474085.3475195 Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W., & McAuley, J. J. (2019). Lakhnes：通过跨领域预训练改进多乐器音乐生成。第20届国际音乐信息检索学会会议论文集，ISMIR 2019，代尔夫特，荷兰，2019年11月4日至8日，685–692。http://archives.ismir.net/ismir201 9/paper/000083.pdf Dong, H.-W., Hsiao, W.-Y., Yang, L.-C., & Yang, Y.-H. (2018). Musegan：用于符号音乐生成与伴奏的多轨序列生成对抗网络。AAAI人工智能会议论文集，32(1)。https://doi.org/10.1609 / aaai.v32i1.11312 Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). 一张图像胜过16×16个词：用于大规模图像识别的 Transfor mer。第9届国际学习表征会议，ICLR 2021，线上会议，奥地利，2021年5月3日至7日。https://openreview.net/forum?id=YicbFdNTTy Eck, D., & Schmidhuber, J. (2002). 使用LSTM循环神经网络进行音乐创作初探。人工智能研究所，103，48。https://peopl e.idsia.ch/~juergen/blues/IDSIA-07-02.pdf Eerola, T., & Vuoskoski, J. K. (2011). 音乐情感离散模型与维度模型的比较。《音乐心理学》，39(1)，18–49。https://doi.org/10.1177/03057356 10362821 Ekman, P. (1971). 情绪面部表情的普遍性与文化差异。内布拉斯加动机研讨会，19，207–283。https://psycnet.apa.org/record/1973-11154-001 Ekman, P. (1975). 揭开面具：识别情绪的面部线索指南。Prentice-Hall。https://psycnet.apa.org/record/1975-31746-000 Engel, J. H., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth：对抗神经音频合成。第7届国际学习表征会议，ICLR 2019，新奥尔良，路易斯安那州，美国，2019年5月6日至9日。https://openreview.net/foru m?id=H1xQVn09FX Ens, J., & Pasquier, P. (2020). 使用多轨音乐机器进行灵活生成。第21届国际音乐信息检索学会会议论文集，ISMIR 2020。https://ddmal.music.mcgill.ca/ISMIR-Conf/static/lbd/ISMIR2020-LBD-423-abstr act.pdf Epps, J., & Holmes, W. H. (1999). 一种用于编码窄带语音宽带增强的新技术。1999年IEEE语音编码研讨会论文集。模型、编码器与误差准则（目录号99EX351），174–176。https://ieeexplore.ieee.org/iel5/6345/16956 /00781522.pdf Evans, Z., Carr, C. J., Taylor, J., Hawley, S. H., & Pons, J. (2024). 快速时间条件潜在音频扩散。第41届国际机器学习会议，ICML 2024，维也纳，奥地利，2024年7月21日至27日。https://openreview.net/forum?id=jOlO8t1xdx Falkowski-Gilski, P., & Uhl, T. (2020). 使用在线流媒体平台消费多媒体内容的当前趋势：以用户为中心的调查。《计算机科学评论》，37，100268。https://doi.org/10.1016/j.cosrev.2020.100268 Fan, Y., Shi, H., Yu, J., Liu, D., Han, W., Yu, H., Wang, Z., Wang, X., & Huang, T. S. (2017). 用于图像超分辨率的平衡式两阶段残差网络。2017年IEEE计算机视觉与模式识别会议

计算机视觉与模式识别研讨会，CVPR 2017研讨会，檀香山，夏威夷州，美国，2017年7月21日至26日，1157–1164。https://doi.org/10.1109/CVPRW.2017.154 Ferreira, L., & Whitehead, J. (2019). 学习生成具有情感的音乐。第20届国际音乐信息检索学会会议论文集，ISMIR 2019，代尔夫特，荷兰，2019年11月4日至8日，384–390。http://archives.ismir.net/ismir201 9/paper/000045.pdf Gan, C., Huang, D., Chen, P., Tenenbaum, J. B., & Torralba, A. (2020). Foley Music：学习从视频生成音乐。计算机视觉—ECCV 2020—第16届欧洲会议，12356，758–775。https://doi.org/10.1007/978-3-030-58621-8_44Gao, X., Zhao, Y., Zhang, J., & Cai, L. (2021). 剧情视频中的成对情感关系识别：数据集与基准。第29届ACM国际多媒体会议论文集，3380–3389。https://doi.org/10.1145/3474085.3475493 Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., & Ritter, M. (2017). 音频集：音频事件的本体与人工标注数据集。2017年IEEE国际声学、语音与信号处理会议，ICASSP 2017，新奥尔良，路易斯安那州，美国，2017年3月5日至9日，776–780。https://doi.org/10.110 9/ICASSP.2017.7952261 Ghodrati, V., S hao, J., Bydder, M., Zhou, Z., Yin, W., Nguyen, K.-L., Yang, Y., & Hu, P. (2019). 使用深度学习进行磁共振图像重建：网络结构和损失函数的评估。《医学与外科定量成像》，9(9)，1516–1527。https://doi.or g/10.21037/qims.2019.08.10 Godsill, S., Rayner, P., & Cappé, O. (2002). 数字音频修复。载于《数字信号处理在音频与声学中的应用》（第133–194页）。Springer US。https://doi.org/10.1 007/0-306-47042-X 4_Gong, Y., Chung, Y.-A., & Glass, J. R. (2021). AST：音频频谱图 Transformer。INTERSPEECH 2021，国际语音通信协会第22届年会，布尔诺，捷克，2021年8月30日至9月3日，571–575。https://doi.org/10.21437/Intersp eech.2021-698 Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.,& Bengio, Y. (2020). 生成对抗网络。《ACM通讯》，63(11)，139–144。https://doi.org/10.1145/3422622 Gunes, H., Schuller, B. W., Pantic, M., & Cowie, R. (2011). 连续空间中的情感表征、分析与合成：一项调查。第九届IEEE自动人脸与手势识别国际会议（FG 2011），圣巴巴拉，加利福尼亚州，美国，2011年3月21日至25日，827–834。https://doi.org/10.1109/FG.2011.5771357 Guo, R., Simpson, I., Magnusson, T., Kiefer, C., & Herremans, D. (2020). 一种由调性张力控制的音乐生成变分自编码器。2020年人工智能音乐创作联合会议论文集。https://arxiv.org/abs/2010.06230 Haidt, J., Rozin, P., Mccauley, C., & Imada, S. (1997). 身体、心理与文化：厌恶与道德之间的关系。《心理学与发展中社会》，9(1)，107–131。http s://doi.org/10.1177/097133369700900105 Harper, F. M., & Konstan, J. A. (2016). MovieLens数据集：历史与背景。《ACM交互式智能系统汇刊》，5(4)，19:1–19:19。https://doi.org/10.1145/2827 872 Haviv, A., Ram, O., Press, O., Izsak, P., & Levy, O. (2022). 无位置编码的 Transformer语言模型仍能学习位置信息。计算语言学协会论文集：EMNLP 2022，1382–1390。https://doi.org/10.18653/V1/2022 .FINDINGS-EMNLP.99

Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., Engel, J. H., Oore, S., & Eck, D.（2018）。起始音与音框：双目标钢琴转录。《第19届国际音乐信息检索学会会议论文集》，50–57。http://ism ir2018.ircam.fr/doc/pdfs/19%5C_Paper.pdf Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. A., Dieleman, S., Elsen, E., Engel, J. H., & Eck, D.（2019）。借助 MAESTRO 数据集实现因子化钢琴音乐建模与生成。《第7届国际学习表征会议》，ICLR 2019，美国路易斯安那州新奥尔良，2019年5月6日至9日。https://openreview.net/forum?id=r1lYRj C9F7 He, K., Zhang, X., Ren, S., & Sun, J.（2016）。用于图像识别的深度残差学习。《IEEE/CVF计算机视觉与模式识别会议》，CVPR 2016，770–778。https://doi.org/10.1109/CVPR.2016.90 Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., Slaney, M., Weiss, R. J., & Wilson, K. W.（2017）。大规模音频分类的卷积神经网络架构。《2017年IEEE国际声学、语音与信号处理会议》，ICASSP 2017，美国路易斯安那州新奥尔良，2017年3月5日至9日，131–135。https://doi.org/10.1109/ICASSP.2017.7952132 Hines, A., Skoglund, J., Kokaram, A. C., & Harte, N.（2012）。ViSQOL：虚拟语音质量客观听评器。《IWAENC 2012—国际声学信号增强研讨会论文集》，德国亚琛工业大学，2012年9月4日至6日。http: // www.vde-verlag.de/proceedings-de/453451035.html Hinton, G., Vinyals, O., & Dean, J.（2015）。提炼神经网络中的知识。arXiv预印本 arXiv:1503.02531。https://doi.org/10.48550/arXiv.1503.02531 Hochreiter, S., & Schmidhuber, J.（1997）。长短期记忆网络。《神经计算》，9(8)，1735–1780。https://doi.org/10.1162/NECO.1997.9.8.1735 Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y.（2020）。神经文本退化的奇特案例。《第8届国际学习表征会议》，ICLR 2020，埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日。https://openreview.net/forum?id=rygGQyrFvH Horn, B. K., & Schunck, B. G.（1981）。确定光流。《人工智能》，17(1-3)，185–203。https://doi.org/10.1016/0004-3702(81)90024-2 Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Q. de Laroussilhe, A. Gesmundo, M. Attariyan, & Gelly, S.（2019）。自然语言处理的参数高效迁移学习。《第36届国际机器学习会议论文集》，ICML 2019，2019年6月9日至15日，美国加利福尼亚州长滩，97，2790–2799。http://proceedings.mlr.press/v97/ho ulsby19a.html Hsiao, W.-Y., Liu, J.-Y., Yeh, Y.-C., & Yang, Y.-H.（2021）。复合词变换器：基于动态有向超图学习整首歌曲音乐的组合。《AAAI人工智能会议论文集》，178–186。https://doi.org/10.1609/AAAI.V35I1.160 91 Hu, E. J., Shen, Y., Wallis, P., Z. Allen-Zhu, Y. Li, Wang, S., Wang, L., & Chen, W.（2022）。LoRA：大语言模型的低秩适配。《第十届国际学习表征会议》，ICLR 2022，线上会议，2022年4月25日至29日。检索日期：2025年2月15日，网址：https://openreview.net/forum?id=nZeVKeeFYf9 Hu, M., & Liu, B.（2004）。挖掘并总结客户评论。《第十届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，美国华盛顿州西雅图，2004年8月22日至25日，168–177。https://doi.org/10.1145/101 4052.1014073

Huang, C. A., Cooijmans, T., Roberts, A., Courville, A. C., & Eck, D.（2017）。卷积对位法。《第18届国际音乐信息检索学会会议论文集》，ISMIR 2017，中国苏州，2017年10月23日至27日，211–218。https: // ismir2017.smcnus.org/wp-content/uploads/2017/10/187%5C_Paper.pdf Huang, C. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D.（2019）。音乐 Transformer：生成具有长期结构的音乐。《第7届国际学习表征会议》，ICLR 2019，美国路易斯安那州新奥尔良，2019年5月6日至9日。https://openreview.net/forum?id=rJe 4ShAcF7 Huang, Q., Xiong, Y., Rao, A., Wang, J., & Lin, D.（2020）。MovieNet：用于电影理解的整体数据集。《计算机视觉—ECCV 2020—第16届欧洲会议》，英国格拉斯哥，2020年8月23日至28日，论文集第四部分，12349，709–727。https://doi.org/10.1007/978-3-030-58548-8 41_Hung, H., Ching, J., Doh, S., Kim, N., Nam, J., & Yang, Y.（2021）。EMOPIA：用于情绪识别和基于情绪的音乐生成的多模态流行钢琴数据集。《第22届国际音乐信息检索学会会议论文集》，ISMIR 2021，在线，2021年11月7日至12日，318–325。https://archives.ismir.net/ismir2021/paper /000039.pdf Hunter, P. G., & Schellenberg, E. G.（2010）。音乐与情感。载于《音乐感知》（第129–164页）。施普林格纽约。https://doi.org/10.1007/978-1-4419-6114-3_5Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., & Kovashka, A.（2017）。自动理解图像和视频广告。《2017年IEEE计算机视觉与模式识别会议》，CVPR 2017，美国夏威夷州檀香山，2017年7月21日至26日，1100–1110。https://doi.org/10.1109/CVPR.2017.123 Ioffe, S., & Szegedy, C.（2015）。批归一化：通过减少内部协变量偏移加速深度网络训练。《第32届国际机器学习会议论文集》，ICML 2015，法国里尔，2015年7月6日至11日，37，448–456。http://proceedings.mlr.press/v37/ioffe15.html Iqbal, T., & Qureshi, S.（2022）。调查：深度学习中的文本生成模型。《沙特国王大学学报—计算机与信息科学》，34(6, Part A)，2515–2528。https://doi.org/10.1016/j.jksuci.2020.04.001 Iser, B., & Schmidt, G.（2003）。用于语音信号带宽扩展的神经网络与码本比较。《第8届欧洲语音通信与技术会议》，EUROSPEECH 2003—INTERSPEECH 2003，瑞士日内瓦，2003年9月1日至4日，565–568。https://doi.org/10.21437/EUROSPEECH.2003-227 Iyyer, M., Enns, P., Boyd-Graber, J. L., & Resnik, P.（2014）。使用递归神经网络检测政治意识形态。《第52届计算语言学协会年会论文集》，ACL 2014，2014年6月22日至27日，美国马里兰州巴尔的摩，第一卷：长篇论文，1113–1122。https://doi.org/10.3115/V1/P14-1105 Izard, C. E.（2013）。人类情感。施普林格科学与商业媒体。http://link.springer.com/10.1007/978-1-4899-2209-0 Jansson, A., Humphrey, E. J., Montecchio, N., Bittner, R. M., Kumar, A., & Weyde, T.（2017）。使用深度U-Net卷积网络进行歌声分离。《第18届国际音乐信息检索学会会议论文集》，ISMIR 2017，中国苏州，2017年10月23日至27日，745–751。https://ismir2017.smcnus.org/wp-content/uploads / 2017/10/171%5C_Paper.pdf Jax, P., & Vary, P.（2003）。电话语音的人工带宽扩展。《信号处理》，83(8)，1707–1719。https://doi.org/10.1016/S0165-1684(03)00082-3

Jhuang, H., Gall, J., Zuffi, S., Schmid, C.与Black, M. J.（2013年）。迈向理解动作识别。IEEE国际计算机视觉会议论文集，3192–3199。检索于2025年2月20日，来自 http://openaccess.thecvf.com/content_icc v_2013/html/Jhuang_Towards_ Understanding_Action_2013_ICCV_paper.html Jiang, Y.-G., Xu, B.与Xue, X.（2014年）。预测用户生成视频中的情绪。AAAI人工智能会议论文集，28(1)。 https://doi.org/10.1609/aaai.v28i 1.8724 Johnson, D. D.（2017年）。使用绑定并行网络生成复调音乐。音乐、声音、艺术与设计中的计算智能，10198，128–143。 https://doi.org/10.1007/9 78-3-319-55750-2\ 9_Juslin, P. N.与Sloboda, J. A.（2001年8月）。音乐与情感：理论与研究。牛津大学出版社。 https://doi.org/10.1093/oso/9780192631886.001.0001 Kang, J., Poria, S.与Herremans, D.（2024）。Video2Music：使用情感多模态Transformer模型从视频生成合适的音乐。《专家系统与应用》，249，123640。https://doi.org/10.1016/j.eswa.2024.123640 Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J.与Amodei, D.（2020年）。神经语言模型的缩放定律。CoRR, abs/2001.08361。https://arxiv.org/abs/2001.08361 Kazemnejad, A., Padhi, I., Natesan Ramamurthy, K., Das, P.与Reddy, S.（2024）。位置编码对Transformer中长度泛化的影响。神经信息处理系统进展，36。https://proceedings.neurips.cc/paper_files/paper/2023/hash/4e85362c02172c0c6567ce593122d31c-Abstract-Conference.html Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C.与Socher, R.（2019年）。CTRL：用于可控生成的条件Transformer语言模型。CoRR，abs/1909.05858。http://ar xiv.org/abs/1909.05858 Keung, P., Lu, Y., Szarvas, G.与Smith, N. A.（2020年）。多语言亚马逊评论语料库。2020年自然语言处理实证方法会议论文集，EMNLP 2020，在线，2020年11月16日至20日，4563–4568。https://doi.org/10.18653 /V1/2020.EMNLP-MAIN.369 Khan, H., Hussain, T., Ullah Khan, S., Ahmad Khan, Z.与Baik, S. W.（2024）。用于监督式视频摘要的深度多尺度金字塔特征网络。《专家系统与应用》，237，121288。 https://doi.org/10.1016/j.eswa.2023.121288 Kilgour, K., Zuluaga, M., Roblek, D.与Sharifi, M.（2019年）。Fréchet 音频距离：一种用于评估音乐增强算法的无参考指标。载于G. Kubin与Z. Kacic（编），国际语音通信协会第20届年会，INTERSPEECH 2019，格拉茨，奥地利，2019年9月15日至19日（第2350–2354页）。ISCA。https://doi . org/10.21437/INTERSPEECH.2019-2219 Kim, J., Lee, S., Kim, S.与Yoo, W. Y.（2011年）。基于唤醒度-效价值的音乐情绪分类模型。第13届先进通信技术国际会议（Icact2011），292–295。https://ieeexplore.ieee.org/iel5/5740523/5745722/05745796.pdf Kim, S.与Sathe, V.（2019年）。通过生成对抗神经网络对原始音频进行带宽扩展。CoRR, abs/1903.09027。http://arxiv.org/abs/1903.09027 Kingma, D. P.与Ba, J.（2015年）。Adam：一种随机优化方法。第三届国际学习表征会议，ICLR 2015，美国加利福尼亚州圣迭戈，2015年5月7日至9日，会议轨道论文集。http://arxiv.org/abs/1412.6980 Kipf, T. N.与Welling, M.（2017年）。使用图卷积神经网络进行半监督分类。第五届国际学习表征会议，ICLR 2017。https://op enreview.net/forum?id=SJU4ayYgl

Koelsch, S.（2014年）。音乐诱发情绪的脑相关性。《自然综述：神经科学》，15(3)，170–180。https://doi.org/10.1038/nrn3666 Koepke, A. S., Wiles, O., Moses, Y.与Zisserman, A.（2020年）。从视觉到声音：一种用于视觉钢琴转录的端到端方法。IEEE国际声学、语音与信号处理会议，ICASSP 2020，1838–1842。https://doi.org/10.1109/ICASSP40776.2020.9053115 Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W.与Plumbley, M. D.（2020年）。PANNs：用于音频模式识别的大规模预训练音频神经网络。IEEE/ACM音频、语音与语言处理汇刊，28，2880–2894。https: / /doi . org/ 10. 1109 /TASLP.2020.3030497 Kontio, J., Laaksonen, L.与Alku, P.（2007）。基于神经网络的语音人工带宽扩展。IEEE音频、语音与语言处理汇刊，15(3)，873–881。https://doi.org/10.1109/TASL.2006.885934 Krishna, K., Wieting, J.与Iyyer, M.（2020年）。将无监督风格迁移重新表述为释义生成。2020年自然语言处理实证方法会议论文集，EMNLP 2020，在线，2020年11月16日至20日，737–762。https://d oi.org/10.18653/V1/2020.EMNLP-MAIN.55 Krizhevsky, A., Sutskever, I.与Hinton, G. E.（2012年）。使用深度卷积神经网络进行ImageNet分类。神经信息处理系统进展25：2012年第26届神经信息处理系统年会。2012年12月3日至6日于美国内华达州太浩湖举行的会议论文集，1106–1114。https: //
p
roceedings.neurips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html Krumhansl, C. L.（2002年）。音乐：认知与情感之间的纽带。《心理科学当前方向》，11(2)，45–50。https://doi.org/10.1111/1467-8721.00165 Kuleshov, V.（2020年四月）。Kuleshov/音频超分辨率。https://github.com/kuleshov/audio-super-r es Kuleshov, V., Enam, S. Z.与Ermon, S.（2017年）。使用神经网络的音频超分辨率。第五届国际学习表征会议，ICLR 2017，法国土伦，2017年4月24日至26日，研讨会轨道论文集。https://openreview.net/forum?id=S1gNakBFx Kusal, S., Patil, S., Choudrie, J., Kotecha, K., Vora, D. R.与Pappas, I. O.（2022年）。基于文本的情绪检测综述——技术、应用、数据集与未来方向。CoRR，abs/2205.03235。https://doi.org/10.48550/ARXIV.2205.03235 Lam, M. W. Y., Tian, Q., Li, T., Yin, Z., Feng, S., Tu, M., Ji, Y., Xia, R., Ma, M., Song, X., Chen, J., Wang, Y.与Wang, Y.（2023年）。高效神经音乐生成。神经信息处理系统进展36：2023年神经信息处理系统年会，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日。http://papers.nips.cc/paper%5C_
files/paper/2023/hash/38b23e2328096520e9c889
ae03e372c9-Abstract-Conference.html Lee, J., Kim, S., Kim, S., Park, J.与Sohn, K.（2019年）。上下文感知情绪识别神经网络。2019年IEEE/CVF 国际计算机视觉会议，ICCV 2019，韩国首尔，2019年10月27日至11月2日，10142–10151。https://doi.org/10.1109 / ICCV.2019.01024 Leon, F. L. D., Wang, Y., Feng, Y.与Lee, M. G.（2025年四月）。Uob-nlp参加SemEval-2025任务11：利用适配器进行多语言与跨语言情绪检测。https://doi.org /10.48550/arXiv.2504.08543 Levesque, H. J., Davis, E.与Morgenstern, L.（2012年）。Winograd模式挑战。知识表示与推理原理：第十三届国际

会议，KR 2012，罗马，意大利，2012年6月10日至14日。
http://www.aaai.org/ocs/index.php /KR/KR12/paper/view/4492 Li, B., Liu, X., Dinesh, K., Duan, Z., & Sharma, G. (2019). 创建用于多模态音乐分析的多轨古典音乐演奏数据集：挑战、见解与应用。《IEEE多媒体汇刊》，21(2), 522–535。
https://doi.org/10.1109/TMM.2 018.2856090 Li, K., Huang, Z., Xu, Y., & Lee, C. (2015). 基于深度神经网络的语音带宽扩展及其在为窄带语音自动语音识别补充高频缺失特征中的应用。第16届国际语音通信协会年会，INTERSPEECH 2015，德国德累斯顿，2015年9月6日至10日，2578–2582。
https://doi.org/10.21437/INTERSPEECH.2015-555 Li, Y., Tagliasacchi, M., Gfeller, B., & Roblek, D. (2020). 学习去除历史音乐中的噪声。第21届国际音乐信息检索学会会议论文集，ISMIR 2020，加拿大蒙特利尔，2020年10月11日至16日，504–511。
http://archives.ismir.net /ismir2020/paper/000057.pdf Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017). 用于单图像超分辨率的增强型深度残差网络。2017年IEEE计算机视觉与模式识别会议研讨会，CVPR 2017研讨会，美国夏威夷州檀香山，2017年7月21日至26日，1132–1140。https://doi.org/10.1109/CVPRW.2017.151 Lim, T., Yeh, R. A.,Xu, Y., Do, M. N., & Hasegawa-Johnson, M. (2018). 用于音频超分辨率的时频网络。2018年IEEE国际声学、语音与信号处理会议，ICASSP 2018，加拿大阿尔伯塔省卡尔加里，2018年4月15日至20日，646–650。https://doi.org/10.1109/ICASSP.2018.8462049 Liu, B., & Zhang, L. (2012). 观点挖掘与情感分析综述。载于《文本数据挖掘》（第415–463页）。施普林格。
https://doi.org/10.1007/978-1-4614-3223-4_13Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y., & Plumb- ley, M. D. (2024). AudioLDM 2：通过自监督预训练学习整体音频生成。《IEEE/ACM音频、语音与语言处理汇刊》，32，2871–2883。https://doi.org/10.1109/TASLP.2024.3399607 Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019年7月). RoBERTa：一种稳健优化的BERT预训练方法。
https://doi.org/10.48550/arXiv.1907.11692 Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). 视频Swin Transformer。IEEE/CVF计算机视觉与模式识别会议，CVPR 2022，3192–3201。https://doi.org/10.1109/CVPR52688.2022.00320 Liutkus, A., Stöter, F., Rafii, Z., Kitamura, D., Rivet, B., Ito, N., Ono, N., & Fontecave, J. (2017). 2016年信号分离评测竞赛。潜变量分析与信号分离——第13届国际会议，LVA/ICA 2017，法国格勒诺布尔，2017年2月21日至23日，论文集，第10169卷，323–332。https://doi.org/10.1007/978-3-319-53547-0 \ 31_Macartney, C., & Weyde, T. (2018). 使用Wave-U-Net改进语音增强。CoRR，abs/1811.11307。
http://arxiv.org/abs/1811.11307 Manocha, P., Finkelstein, A., Zhang, R., Bryan, N. J., Mysore, G. J., & Jin, Z. (2020). 基于最小可觉差学习的可微感知音频指标。国际语音通信协会第21届年会，Interspeech 2020，线上会议，中国上海，2020年10月25日至29日，2852–2856。https://doi.org/10.214 37/INTERSPEECH.2020-1191

Matsuda, Y.-T., Fujimura, T., Katahira, K., Okada, M., Ueno, K., Cheng, K., & Okanoya, K. (2013). 类别型与维度策略的内隐加工：一项关于面部情绪知觉的功能性磁共振成像研究。《人类神经科学前沿》，第7卷，2013年。https: //doi.org/10.3389/fnhum.2013.00551 May, P. (2021). 机器翻译多语言STS基准数据集。https://github.com/Philip May/stsb-multi-mt McFee, B., Humphrey, E. J., & Bello, J. P. (2015). 用于音乐数据增强的软件框架。第16届国际音乐信息检索学会会议论文集，ISMIR 2015，西班牙马拉加，2015年10月26日至30日，248–254。http://ismir201 5.uma.es/articles/228%5C_Paper.pdf Meyer, L. B. (2008). 音乐中的情绪与意义。芝加哥大学出版社。https://press.u chicago.edu/ucp/books/book/chicago/E/bo28551887.html Mikolov, T., & Zweig, G. (2012). 上下文相关循环神经网络语言模型。2012年IEEE口语语言技术研讨会（SLT），美国佛罗里达州迈阿密，2012年12月2日至5日，234–239。https://doi.org/10.1109/SLT.2012.6424228 Miranda, E. R., Magee, W. L., Wilson, J. J., Eaton, J., & Palaniappan, R. (2011). 脑机音乐接口（BCMI）：从基础研究迈向特殊需求群体的现实应用。《音乐与医学》，3(3), 134–140。https://doi.org/10.47513/mmd.v3i3.370 Miron, M., & Davies, M. E. P. (2018). 使用卷积自编码器重建音乐混合音频的高频幅度谱图。第21届国际数字音频效果会议（DAFx2018），173–180。https://www.researchgate.net/publication/329372321_High_frequency_magnitude_spectrogram_reconstruction_for_music_mixture s_using_convolutional_autoencoders Miyazawa, K., Kyuragi, Y., & Nagai, T. (2022). 一种简单有效的基于预训练Transformer模型的多模态学习方法。《IEEE Access》，10，29821–29833。https://doi.org/10.11 09/ACCESS.2022.3159346 Mohammad, S. M., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018任务1：推文中的情感。第12届语义评测国际研讨会论文集，SemEval@NAACL-HLT 2018，美国路易斯安那州新奥尔良，2018年6月5日至6日，1–17。https://doi.org/10.18653/V1/S18-1001 Mokady, R., Hertz, A., & Bermano, A. H. (2021). CLIPCap：用于图像描述的CLIP前缀。CoRR, abs/2111.09734。https://arxiv.org/abs/2111.09734 Morgan, S. D. (2019). 情感语音的类别型与维度评分：Morgan情感语音集的行为研究结果。《言语、语言与听力研究杂志》，62(11), 4015–4029。https://doi.org/10.1044/2019_JSLHR-S-19-0144 Muhammad, S. H., Ousidhoum, N., Abdulmumin, I., Yimam, S. M., Wahle, J. P., Ruas, T., Be- loucif, M., de Kock, C., Belay, T. D., Ahmad, I. S., Surange, N., Teodorescu, D., Adelani, D. I., Aji, A. F., Ali, F., Araujo, V., Ayele, A. A., Ignat, O., Panchenko, A., ... Mo- hammad, S. M. (2025). SemEval-2025任务11：弥合基于文本的情绪检测鸿沟。CoRR，abs/2503.07269。https://doi.org/10.48550/ARXIV.2503.07269 Müller, M. (2015). 音乐处理基础：音频、分析、算法与应用。施普林格国际出版公司。https://doi.org/10.1007/978-3-319-21945-5 Nair, V., & Hinton, G. E. (2010). 修正线性单元改进受限玻尔兹曼机。第27届国际机器学习会议（ICML-10）论文集，807–814。检索日期：2025年2月22日，网址：https://www.cs.toronto.edu/~hinton/absps/relu ICML.pdf Nakatoh, Y., Tsushima, M., & Norimatsu, T. (1997). 使用分段线性映射从窄带语音生成宽带语音。第五届欧洲语音会议

《通信与技术》，EUROSPEECH 1997，希腊罗得岛，1997年9月22日至25日，1643–1646。https://doi.org/10.21437/EUROSPEECH.1997-469 Nguyen, T. H., Shirai, K., & Velcin, J. (2015). 面向股票走势预测的社交媒体情感分析。《专家系统与应用》，42(24)，9603–9611。https://doi.org/10 .1016/J.ESWA.2015.07.052 Nitanda, N., Haseyama, M., & Kitajima, H. (2005). 用于场景切换检测的音频信号分割与分类。电路与系统国际研讨会（ISCAS 2005），2005年5月23日至26日，日本神户，4030–4033。https://doi.org/10.1109/ISCAS.2005.1465515 Niu, S., Liu, Y., Wang, J., & Song, H. (2020). 迁移学习十年综述（2010–2020）。IEEE人工智能汇刊，1(2)，151–166。https: //doi .org/ 10.1109 /TAI.2021.3054609 Nour-Eldin, A. H., & Kabal, P. (2008). 基于梅尔频率倒谱系数的窄带语音带宽扩展。国际语音通信协会第九届年会INTERSPEECH 2008，澳大利亚布里斯班，2008年9月22日至26日，53–56。https://doi.org/10.21437/INTERSPEECH.2008-11 Ojala, T., Pietikäinen, M., & Harwood, D. (1996). 基于特征分布的分类与纹理测度比较研究。《模式识别》，29(1)，51–59。https://d oi.org/10.1016/0031-3203(95)00067-4 Ojha, A. K., Doˇgruöz, A. S., Madabushi, H. T., Martino, G. D. S., Rosenthal, S., & Rosá, A. (编). (2024). 第18届语义评测国际研讨会论文集，SemEval@NAACL 2024，墨西哥墨西哥城，2024年6月20日至21日。计算语言学协会。https://aclanthology.org/volumes/2024.semeval-1/ Oore, S., Simon, I., Dieleman, S., Eck, D., & Simonyan, K. (2020). 这一次，融入情感：表现性音乐演奏学习。《神经计算与应用》，32(4)，955–967。https://doi.org/10.1007/S00521-018-3758-9 Ovalle, J. E. A., Solorio, T., Montes-y-Gómez, M., & González, F. A. (2017). 用于信息融合的门控多模态单元。第5届国际学习表征会议，ICLR 2017，法国土伦，2017年4月24日至26日，研讨会论文集。检索于2025年2月24日，网址：https://openreview.net/forum?id=S12%5C_nquOe Pan, J., Wang, S., & Fang, L. (2022). 通过多模态注意力和时间同步评论进行表征学习，以分析情感视频内容。MM ’22：第30届ACM多媒体国际会议，葡萄牙里斯本，2022年10月10日至14日，42–50。https://doi.org/10.1145/3503161.3548018 Panda, R., Malheiro, R., Rocha, B., Oliveira, A., & Paiva, R. P. (2013). 多模态音乐情感识别：一种新数据集、方法论与比较分析。计算机音乐多学科研究国际研讨会。https://hdl.handle.net/10316 /94095 Pandeya, Y. R., & Lee, J. (2021). 基于深度学习的多模态信息后期融合，用于音乐视频情绪分类。《多媒体工具与应用》，80(2)，2887–2905。https://doi.org/10.1007/s11042-020-08836-3 Pang, B., Lee, L., & Vaithyanathan, S. (2002). 点赞？使用机器学习技术进行情感分类。2002年自然语言处理实证方法会议论文集，EMNLP 2002，美国宾夕法尼亚州费城，2002年7月6日至7日，79–86。https://doi.org/10.3115/1118693.1118704 Park, K., & Kim, H. S. (2000a). 使用基于高斯混合模型的变换进行语音窄带到宽带转换。IEEE声学、语音与信号处理国际会议。ICASSP 2000，2000年6月5日至9日，土耳其伊斯坦布尔希尔顿酒店及会议中心，1843–1846。https://doi.org/10.1109/ICASSP.2000.862114

Park, K., & Kim, H. S. (2000b). 使用基于高斯混合模型的变换进行语音窄带到宽带转换。IEEE声学、语音与信号处理国际会议。ICASSP 2000，2000年6月5日至9日，土耳其伊斯坦布尔希尔顿酒店及会议中心，1843–1846。https://doi.org/10.1109/ICASSP.2000.862114 Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch：一种命令式风格、高性能的深度学习库。神经信息处理系统进展32：神经信息处理系统年会2019，NeurIPS 2019，2019年12月8日至14日，加拿大不列颠哥伦比亚省温哥华，8024–8035。https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012 727740-Paper.pdf Payne, C. (2019). MuseNet。OpenAI博客，3。https://openai.com/blog/musenet Perraudin, N., Holighaus, N., Majdak, P., & Balazs, P. (2018). 利用相似性图修复长音频片段。IEEE/ACM 音频、语音与语言处理汇刊，26(6)，1079–1090。https://doi.org/10.1109/TASLP.2018.2809864 Pimpale, M. M. R., Therese, S., & Shinde, V. (2016). 声源分离方法综述。《国际期刊》，3(11)，580–584。https://www.academia.edu/download/50739 135/V3I1103.pdf Plutchik, R. (1988年). 情绪的本质：临床意义。载于《情绪与精神病理学》（第1–20页）。Springer US。https://doi.org/10.1007/978-1-4757-1987-1_1Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2019). MELD：用于对话情绪识别的多模态多方数据集。第57届计算语言学协会会议论文集，ACL 2019，意大利佛罗伦萨，2019年7月28日至8月2日，第1卷：长篇论文，527–536。https://doi.org/10 .18653/V1/P19-1050 Pratt, R. R., Abel, H.-H., & Skidmore, J. (1995). 背景音乐辅助神经反馈训练对ADD和ADHD儿童脑电图模式的影响。《国际艺术医学杂志》，4(1)，24–31。https://psycnet.apa.org/record/1996-02780-004 Prendergast, R. M. (1992). 电影音乐：被忽视的艺术——电影音乐批评研究。WW Norton & Company。https://books.google.com/books/about/Film_Music.html?id=7ozTu s2HZbsC Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). 通过自然语言监督学习可迁移的视觉模型。第38届国际机器学习会议论文集，ICML 2021，2021年7月18日至24日，线上会议，139，8748–8763。http://proceedings.mlr.press/v139/radford21a.html Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). 通过大规模弱监督实现鲁棒语音识别。国际机器学习会议，ICML 2023，2023年7月23日至29日，美国夏威夷州檀香山，202，28492–28518。https://proceedings.mlr.press/v202/radford23a.html Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). 通过生成式预训练提升语言理解能力。OpenAI。http://www.nlpir.org/wordpress/wp-content/uploads/2019/06/Improving-language-understanding-by-generative-pre-training.pdf Raffel, C. (2016). 基于学习的方法：比较序列及其在音频到MIDI对齐与匹配中的应用 [博士论文，哥伦比亚大学，美国]。https://doi.org/10.7916/D8N58MHV

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). 使用统一的文本到文本 Transformer 探索迁移学习的极限。《机器学习研究杂志》，21(1)。Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2017). 用于音乐分离的 MUSDB18 语料库。 https://inria.hal.science/hal-02190845/document Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., & Bittner, R. (2019, August). MUSDB18-HQ——MUSDB18 的无压缩版本。 https://doi.org/10.5281/zenodo.3338372 Ranjbar, N., & Baghbani, H. (2025, April). Lotus 在 SemEval-2025 任务11中的方案：使用 Llama-3 生成的解释与 RoBERTa 进行多标签情绪分类。https://doi.org/10.48550/ar Xiv.2502.19935 Ray, A., & Kolekar, M. H. (2024). 迁移学习及其在人类活动识别中的广泛适用性：一项调查。《专家系统与应用》，240, 122538。https://do i.org/10.1016/j.eswa.2023.122538 Redmon, J., Divvala, S. K., Girshick, R. B., & Farhadi, A. (2016). You Only Look Once：统一的实时目标检测。2016年IEEE计算机视觉与模式识别会议，CVPR 2016，美国内华达州拉斯维加斯，2016年6月27日至30日，779–788。https://doi.org/10.1 109/CVPR.2016.91 Reiss, T., Cohen, N., Bergman, L., & Hoshen, Y. (2021). PANDA：调整预训练特征以用于异常检测和分割。IEEE计算机视觉与模式识别会议，CVPR 2021，线上，2021年6月19日至25日，2806–2814。https://doi.org/10.1109 /CVPR46437.2021.00283 Rempfler, M., Kumar, S., Stierle, V., Paulitschke, P., Andres, B., & Menze, B. H. (2017). 无透镜显微镜视频中的细胞谱系追踪。医学图像计算与计算机辅助干预——MICCAI 2017——第20届国际会议，魁北克市，加拿大魁北克省，2017年9月11日至13日，论文集，第二部分，10434，3–11。https://doi.org/10.10 07/978-3-319-66185-8\ 1_Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). 语音质量感知评估（ PESQ ） —— 一种用于评估电话网络和编解码器语音质量的新方法。IEEE声学、语音与信号处理国际会议，ICASSP 2001，2001年5月7日至11日，美国犹他州盐湖城盐宫会议中心，论文集，749–752。https://doi.org/10.1109/ICASSP.2001.941023 Roberts, A., Engel, J. H., Raffel, C., Hawthorne, C., & Eck, D. (2018). 用于学习音乐长期结构的分层潜向量模型。第35届国际机器学习会议论文集，ICML 2018，瑞典斯德哥尔摩斯德哥尔摩国际展览中心，2018年7月10日至15日，80，4361–4370。http://proceedings.mlr.press/v80/roberts18a.html Rodriguez Bribiesca, I., Lopez Monroy, A. P., & Montes-y-Gomez, M. (2021). 用于电影类型分类的 Transformer 多模态加权融合。第三届多模态人工智能研讨会论文集，1–5。https://doi.org/10.18653/v1/2021.m aiworkshop-1.1 Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net：用于生物医学图像分割的卷积网络。医学图像计算与计算机辅助干预——MICCAI 2015——第18届国际会议，德国慕尼黑，2015年10月5日至9日，论文集，第三部分，9351，234–241。https://doi.org/10.1007/978-3-319-24574-4_28Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). 通过反向传播误差学习表征。《自然》，323(6088)，533–536。https://doi.org/10.1038/323533a0 Russell, J. A. (1980). 情感环形模型。《人格与社会心理学杂志》，39(6)，1161。https://doi.org/10.1037/h0077714

Russell, J. A., & Mehrabian, A. (1977). 情绪三因素理论的证据。《人格研究杂志》，11(3)，273–294。https://doi.org/https://doi.org/10.1016/0092-65 66(77)90037-X Sambaragi, L. M., Pradeep Naik, P., Kakatkar, N. N., Chikkamath, S., S R, N., & Budihal, S. V. (2024). 音乐生成：一种简化方法。2024年第三届技术创新国际会议（INOCON），1–6。https://doi.org/10.1109/INOCON60754.2024 .10511477 Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT，BERT 的蒸馏版本：更小、更快、更便宜、更轻量。arXiv 预印本 arXiv:1910.01108。https://doi.org/10 . 48550/arXiv.1910.01108 Sautter, J., Faubel, F., Buck, M., & Schmidt, G. (2019). 使用结合判别式训练的条件生成对抗网络进行人工带宽扩展。IEEE国际声学、语音与信号处理会议，ICASSP 2019，英国布莱顿，2019年5月12日至17日，7005–7009。https://doi.org/10.1109/ICASSP.2019.8682649 Savchenko, A. V. (2023). Emotieffnets：用于基于视频的效价-唤醒度预测、表情分类和动作单元检测的面部处理。IEEE/CVF计算机视觉与模式识别会议，CVPR 2023——研讨会，加拿大不列颠哥伦比亚省温哥华，2023年6月17日至24日，5716–5724。https://doi.org/10.1109/CVPRW59228.2023.00606 Schifrin, L. (2011). 电影与电视音乐创作。Hal Leonard Corporation。https://b ooks.google.com/books/about/Music_ Composition_for_Film_and_Televisio.html?id=g TwLAQAAQBAJ Sennrich, R., Haddow, B., & Birch, A. (2016a). 通过侧向约束控制神经机器翻译中的礼貌表达。NAACL HLT 2016，2016年计算语言学协会北美分会人类语言技术会议，美国加利福尼亚州圣迭戈，2016年6月12日至17日，35–40。https://doi.org/10.18653/v1/n 16-1005 Sennrich, R., Haddow, B., & Birch, A. (2016b). 使用子词单元进行罕见词的神经机器翻译。计算语言学协会第54届年会论文集，ACL 2016，2016年8月7日至12日，德国柏林，第一卷：长篇论文。https://doi.org/10.18653/V1/P16-1162 Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). 机器学习三个时代的计算趋势。国际神经网络联合会议，IJCNN 2022，意大利帕多瓦，2022年7月18日至23日，1–8。https://doi.org/10.1109/IJCNN55064.2 022.9891914 Shahraki, A. G. (2015). 文本情绪挖掘 [硕士论文，阿尔伯塔大学]。https: // era.library.ualberta.ca/items/27ae961f-d9a6-4a5a-9b6f-f180478ea573 Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). 具有相对位置表示的自注意力。2018年计算语言学协会北美分会人类语言技术会议论文集，NAACL-HLT，美国路易斯安那州新奥尔良，2018年6月1日至6日，第二卷（短篇论文），464–468。https://doi.org/10.18 653/V1/N18-2074 Sheng, E., Chang, K.-W., Natarajan, P., & Peng, N. (2020). 朝着可控的语言生成偏见迈进。2020年自然语言处理经验方法会议论文集：研究结果，EMNLP 2020，线上活动，2020年11月16日至20日，EMNLP 2020，3239–3254。https://doi.org/10.18653/v1/2020.findings-emnlp.291 Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., & Wang, Z. (2016). 使用高效的亚像素卷积神经网络实现单幅图像和视频的实时超分辨率。2016年IEEE计算机视觉与模式

识别，CVPR 2016，美国内华达州拉斯维加斯，2016年6月27日至30日，1874–1883。https://doi .org/10.1109/CVPR.2016.207 Simoes, G. S., Wehrmann, J., Barros, R. C., & Ruiz, D. D.（2016）。使用卷积神经网络进行电影类型分类。2016年国际神经网络联合会议，IJCNN 2016，加拿大不列颠哥伦比亚省温哥华，2016年7月24日至29日，259–266。https://doi.org/10.110 9/IJCNN.2016.7727207 Simonyan, K., & Zisserman, A.（2015）。用于大规模图像识别的极深卷积网络。http://arxiv.org/abs/1409.1556 Smith, E. M., Gonzalez-Rico, D., Dinan, E., & Boureau, Y.（2020）。控制生成对话中的风格。CoRR，abs/2009.10855。https://arxiv.org/abs/2009.10855 Sohl-Dickstein, J., Weiss, E. A., Maheswaranathan, N., & Ganguli, S.（2015）。利用非平衡热力学进行深度无监督学习。第32届国际机器学习会议论文集，ICML 2015，法国里尔，2015年7月6日至11日，37，2256–2265。http://proceedings.mlr.press/v37/sohl-dickstein15.html Soleymani, M., Pantic, M., & Pun, T.（2012）。对视频作出反应时的多模态情感识别。IEEE 情感计算汇刊，3(2)，211–223。https://doi.org/10.110 9/T-AFFC.2011.37 Song, G.-B., & Martynovich, P.（2009）。基于隐马尔可夫模型的语音信号带宽扩展研究。《信号处理》，89(10)，2036–2044。https://doi.org/10.1016/j.sigpro.2009.03 .037 Srivastava, N., G. Hinton, Krizhevsky, A., Sutskever, I., & Salakhutdinov, R.（2014）。随机失活：一种防止神经网络过拟合的简便方法。《机器学习研究杂志》，15(1)，1929–1958。https://doi.org/10.5555/2627435.2670313 Stratton, V. N., & Zalanowski, A. H.（1994）。音乐与歌词的情感影响。《艺术实证研究》，12(2)，173–184。https://doi.org/10.2190/35T0-U4DT-N09Q-LQHW Su, K., Liu, X., & Shlizerman, E.（2020）。Audeo：为无声表演视频生成音频。神经信息处理系统进展 33：2020年神经信息处理系统年会，NeurIPS 2020，2020年12月6日至12日，线上举行。https://proceedings.neurips.cc/paper/2020/hash/227f6afd3b7f89b96c4bb91f95d50f6d-Abstract.html Su, K., Liu, X., & Shlizerman, E.（2021）。听起来如何？神经信息处理系统进展 34：2021年神经信息处理系统年会，NeurIPS 2021，2021年12月6日至14日，线上举行，29258–29273。https://proceedings.neurips.cc/paper/2021/hash/f4e369c0a468d3aeeda0593ba90b5e55-Abstract.html Sulun, S.（2018）。用于视频压缩的深度学习帧预测。arXiv 预印本 arXiv:1811.10946。http://arxiv.org/abs/1811.10946 Sulun, S., & Davies, M. E. P.（2021）。关于使用深度神经网络进行音乐带宽扩展时的滤波器泛化。IEEE 信号处理选题期刊，15(1)，132–142。https://doi.org/10.1109/JSTSP.2020.3037485 Sulun, S., Davies, M. E. P., & Viana, P.（2022）。基于连续值情绪条件的符号音乐生成。IEEE Access，10，44617–44626。https://doi.org /10.1109 /ACCESS.2022.3169744 Sulun, S., Oliveira, P., & Viana, P.（2023）。Emotion4MIDI：基于歌词、带有情绪标签的符号音乐数据集。Progress in Artificial Intelligence——第22届EPIA人工智能会议，EPIA 2023，亚速尔群岛法亚尔岛，2023年9月5日至8日，论文集第二部分，14116，77–89。https://doi.org/10.1007/978-3-031-49011-8_7

Sulun, S., & Tekalp, A. M.（2021）。学习式帧预测能否在视频编码方面与块运动补偿媲美？《信号、图像与视频处理》，15(2)，401–410。https://doi .org/10.1007/S11760-020-01751-Y Sulun, S., Viana, P., & Davies, M. E. P.（2024a）。使用多模态预训练特征进行电影预告片类型分类。《专家系统与应用》，125209。https://doi.org/10.1 016/j.eswa.2024.125209 Sulun, S., Viana, P., & Davies, M. E. P.（2024b）。VEMOCLAP：视频情感分类网页应用。2024年国际多媒体研讨会（ISM），137–140。https://d oi.org/10.1109/ISM63611.2024.00029 Sulun, S., Viana, P., & Davies, M. E. P.（2026）。通过对齐情绪与时间边界生成视频配乐 [即将刊印]。《IEEE多媒体汇刊》。https://arxiv.org/a bs/2502.10154 Sun, D. L., & Mazumder, R.（2013）。用于带宽扩展的非负矩阵补全：一种凸优化方法。IEEE信号处理机器学习国际研讨会，MLSP 2013，英国南安普敦，2013年9月22日至25日，1–6。https://doi.org/10.1109/MLSP.2013.6661924 Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C.（2018）。深度迁移学习综述。载于《人工神经网络与机器学习——ICANN 2018》（第270–279页，第11141卷）。施普林格国际出版。https://doi.org/10.1007/978-3-030-01424 -7 27_Tan, H. H., & Herremans, D.（2020）。Music fadernets：通过低层特征建模，基于高层特征实现可控音乐生成。第21届国际音乐信息检索学会会议论文集，ISMIR 2020，加拿大蒙特利尔，2020年10月11日至16日，109–116。http://archives.ismir.net/ismir2020/paper/000222.pdf Tan, M., & Le, Q.（2019）。EfficientNet：重新思考卷积神经网络的模型缩放。第36届国际机器学习会议论文集，97，6105–6114。https://proceedings.mlr.press/v97/tan19a.html Taylor, L., & Nitschke, G.（2018）。利用通用数据增强改进深度学习。IEEE计算智能研讨会系列，SSCI 2018，印度班加罗尔，2018年11月18日至21日，1542–1547。https://doi.org/10.1109/SSCI.2018.8628742 Thao, H. T. P., Roig, G., & Herremans, D.（2023）。Emomv：用于分类和检索的情感性音乐-视频对应学习数据集。《信息融合》，91，64–79。https://d oi.org/10.1016/j.inffus.2022.10.002 Torcoli, M., Wu, C.-W., Dick, S., Williams, P. A., Modar Halimeh, M., Wolcott, W., & Habets, E. A. P.（2024）。ODAQ：开放音频质量数据集。ICASSP 2024——2024年IEEE声学、语音与信号处理国际会议（ICASSP），836–840。https://doi.org/10.1109/ICASSP48485.2024.10447634 Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I.（2017）。注意力就是你所需要的一切。神经信息处理系统进展 30：2017年神经信息处理系统年会，2017年12月4日至9日，美国加利福尼亚州长滩，5998–6008。https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html Vinyals, O., Toshev, A., Bengio, S., & Erhan, D.（2017）。Show and tell：从2015年MSCOCO图像描述挑战赛中汲取的经验。IEEE模式分析与机器智能汇刊，39(4)，652–663。https://doi.org/10.1109/TPAMI.2016.2587640

Viveros-Muñoz, R., Huijse, P., Vargas, V., Espejo, D., Poblete, V., Arenas, J. P., Vernier, M., Ver- gara, D., & Suárez, E. (2023). 用于城市声景复音声音事件检测任务的数据集：合成复音环境声源（SPASS）数据集. Data in Brief, 50, 109552. https://doi.org/10.1016/j.dib.2023.109552 Wang, B., Zhao, D., Lioma, C., Li, Q., Zhang, P., & Simonsen, J. G. (2020). 在复数嵌入中编码词序. 第8届国际学习表征会议，ICLR 2020，埃塞俄比亚亚的斯亚贝巴，2020年4月26日至30日. https://openreview.net/forum?id=Hke- WTVtwr Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Gool, L. V. (2016). 时间片段网络：迈向深度动作识别的良好实践. 计算机视觉—ECCV 2016—第14届欧洲会议，荷兰阿姆斯特丹，2016年10月11日至14日，论文集，第VIII部分，9912，20–36. https://doi.org/10.1007/978-3-319-46484-8_2Wang, L., Zhang, X., Su, H., & Zhu, J. (2024). 持续学习综述：理论、方法与应用. IEEE Transactions on Pattern Analysis Machine Intelli- gence, 46(8), 5362–5383. https://doi.org/10.1109/TPAMI.2024.3367329 Wang, Z., Wang, D., Zhang, Y., & Xia, G. (2020). 学习可控复调音乐生成的可解释表征. 第21届国际音乐信息检索学会会议论文集，ISMIR 2020，加拿大蒙特利尔，2020年10月11日至16日，662–669. http://archives.ismir.net/ismir2020/paper/000094.pdf Wehrmann, J., & Barros, R. C. (2017). 电影类型分类：一种基于时间卷积的多标签方法. Applied Soft Computing, 61, 973–982. https://doi.org/10.1 016/j.asoc.2017.08.029 Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). 思维链提示激发大语言模型的推理能力. http: //papers.nips.cc/paper%5C_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html Wei, J., Yang, X., & Dong, Y. (2021). 基于关键帧的用户生成视频情感识别. 多媒体工具与应用, 80(9), 14343–14361. https://doi.org/10.1007 / S11042-020-10203-1 Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). 迁移学习综述. Journal of Big data, 3, 1–40. https://doi.org/10.1186/S40537-016-0043-6 Weston, J., Dinan, E., & Miller, A. H. (2018). 检索与优化：改进对话序列生成模型. 第2届面向搜索的对话式人工智能国际研讨会论文集，SCAI@EMNLP 2018，比利时布鲁塞尔，2018年10月31日，87–92. https://doi.org/10.18653/v1/w18-5713 Whitelaw, C., Garg, N., &Argamon,S. (2005). 使用评价组进行情感分析. 2005年ACM CIKM国际信息与知识管理会议论文集，德国不来梅，2005年10月31日至11月5日，625–631. https: // doi.org/10.1145/1099554.1099714 Wiem, M. B. H., & Lachiri, Z. (2017). 使用MAHNOB-HCI数据库在唤醒度—效价模型中进行情绪分类. International Journal of Advanced Computer Science and Applica- tions, 8(3). https://doi.org/10.14569/IJACSA.2017.080344 Williams, D., Kirke, A., Eaton, J., Miranda, E., Daly, I., Hallowell, J., Roesch, E., Hwang, F., & Nasuto, S. J. (2015). 根据持续变化的情绪轨迹生成动态游戏配乐. 音频工程学会会议：第56届国际会议：游戏音频. https://aes2.org/publications/elibrary-page/?id=17593

Williams, D., Kirke, A., Miranda, E. R., Roesch, E., Daly, I., & Nasuto, S. (2015). 探究算法作曲系统中的情感. 《音乐心理学》, 43(6), 831–854. https: // doi.org/10.1177/0305735614543282 Wingstedt, J., Berg, J., Liljedahl, M., & Lindberg, S. (2005). Remupp：一种用于评估音乐参数与感知属性之间关系的界面. 计算机娱乐技术进展国际会议论文集，ACE 2005，西班牙瓦伦西亚，2005年6月15日至15日，346–349. https://doi.org/10.1145/1178477.1178543 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., & Brew, J. (2019). Huggingface 的 Transformers：最先进的自然语言处理. CoRR, abs/1910.03771. http://arxiv.org/abs/1910.03771 Won, M., Chun, S., Nieto, O., & Serrc, X. (2020). 用于音频表征学习的数据驱动谐波滤波器. ICASSP 2020—2020 IEEE国际声学、语音与信号处理会议（ICASSP），536–540. https://doi.org/10.1109/ICASSP40776.2020 .9053669 Wu, S., Xu, J., Tai, Y., & Tang, C. (2018). 具有大幅前景运动的深度高动态范围成像. 计算机视觉—ECCV 2018—第15届欧洲会议，德国慕尼黑，2018年9月8日至14日，论文集，第II部分，11206，120–135. https://doi.org/10.1007/97 8-3-030-01216-8\ 8_Wu, S., & Yang, Y. (2020). 爵士乐 Transformer 前沿探索：通过定量指标探究人工智能作曲音乐的不足. 第21届国际音乐信息检索学会会议论文集，ISMIR 2020，142–149. http://archiv es.ismir.net/ismir2020/paper/000339.pdf Xu, B., Fu, Y., Jiang, Y.-G., Li, B., & Sigal, L. (2018). 视频情感识别、归因与摘要生成中的异构知识迁移. IEEE TransactionsonAffective Computing, 255–270. https://doi.org/10.1109/TAFFC.2016.2622690 Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., & Bengio, Y. (2015). 展示、关注与讲述：具有视觉注意力的神经图像描述生成. 第32届国际机器学习会议论文集，ICML 2015，法国里尔，2015年7月6日至11日，37，2048–2057. http://proceedings.mlr.press/v37/xuc15.ht ml Xu, Q., Huang, G., Yuan, Y., Guo, C., Sun, Y., Wu, F., & Weinberger, K. Q. (2018). 生成对抗网络评估指标的实证研究. CoRR, abs/1806.07755. http://arxiv.org/abs/1806.07755 Yadav, A., & Vishwakarma, D. K. (2020). 用于电影预告片类型分类的深度网络统一框架. Applied Soft Computing, 96, 106624. https://doi.org/10.101 6/j.asoc.2020.106624 Yang, L., Chou, S., & Yang, Y. (2017). MidiNet：用于符号域音乐生成的卷积生成对抗网络. 第18届国际音乐信息检索学会会议论文集，ISMIR 2017，中国苏州，2017年10月23日至27日，324–331. https://ismir2017.smcnus.org/wp-content/uploads/2017/10/226%5C_Pa per.pdf Yang, R., Wang, D., Wang, Z., Chen, T., Jiang, J., & Xia, G. (2019). 通过潜在表征解耦实现深度音乐类比. 第20届国际音乐信息检索学会会议论文集，ISMIR 2019，荷兰代尔夫特，2019年11月4日至8日，596–603. http://archives.ismir.net/ismir2019/paper/000072.pdf Yang, Y., Lin, Y., Cheng, H. T., Liao, I., Ho, Y., & Chen, H. H. (2008). 迈向多模态音乐情感分类. 收录于《多媒体信息处理进展—PCM 2008》（第5353卷，第70–79页）. 施普林格. https://doi.org/10.1007/978-3-540-89796-5_8

Yariv, G., Gat, I., Benaim, S., Wolf, L., Schwartz, I., & Adi, Y. (2024). 通过文本到视频模型适配实现多样且对齐的音频到视频生成。《AAAI人工智能会议论文集》，38，6639–6647。检索日期：2025年6月20日，网址：https://ojs . aaai.org/index.php/AAAI/article/view/28486 Yi, Y., Zhou, J., Wang, H., Tang, P., & Wang, M. (2024). 基于长程相关性感知网络的用户生成视频情绪识别。《IET图像处理》，18(12)，3288–3301。https://doi.org/10.1049/IPR2.13174 Yoshida, Y., & Abe, M. (1994). 一种基于码本映射、从窄带语音重建宽带语音的算法。第三届国际口语语言处理会议，ICSLP 1994，日本横滨，1994年9月18日至22日，1591–1594。https://doi.org/10.21437/ICSLP.1994-412 Yu, Z., Yu, J., Fan, J., & Tao, D. (2017). 用于视觉问答的多模态因子化双线性池化与协同注意力学习。IEEE国际计算机视觉会议，ICCV 2017，意大利威尼斯，2017年10月22日至29日，1839–1848。https://doi.org/10.1109 /ICCV.2017.202 Zagoruyko, S., & Komodakis, N. (2016). 宽残差网络。2016年英国机器视觉会议论文集，BMVC 2016，英国约克，2016年9月19日至22日。https://b mva-archive.org.uk/bmvc/2016/papers/paper087/index.html Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). 对话式问答：一项调查。《知识与信息系统》，64(12)，3151–3195。https://d oi.org/10.1007/s10115-022-01744-y Zander, M. F. (2006). 广告中的音乐影响：音乐如何改变人们对产品代言人和品牌的第一印象。《音乐心理学》，34(4)，465–480。https://doi.org/10.1 177/0305735606067158 Zhang, H., & Xu, M. (2023). 通过帧级适配和情绪强度学习识别用户生成视频中的情绪。《IEEE多媒体汇刊》，25，881–891。https://doi.org/10.1109/TMM.2021.3134167 Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F., & Wang, G. (2023). 大型语言模型的指令微调：一项调查。CoRR，abs/2308.10792。https://doi.org/10.48550/ARXIV.2308.10792 Zhang, Y., Tian, Y., Kong, Y., Zhong, B., & Fu, Y. (2018). 用于图像超分辨率的残差密集网络。2018年IEEE计算机视觉与模式识别会议，CVPR 2018，美国犹他州盐湖城，2018年6月18日至22日，2472–2481。https://doi.org/10.1 109/CVPR.2018.00262 Zhang, Z., Wang, L., & Yang, J. (2023). 通过跨模态时间擦除网络进行弱监督视频情绪检测与预测。IEEE/CVF计算机视觉与模式识别会议，CVPR 2023，加拿大不列颠哥伦比亚省温哥华，2023年6月17日至24日，18888–18897。https://doi.org/10.1109/CVPR52729.2023.01811 Zhang, Z., Zhao, P., Park, E., & Yang, J. (2024). MART：通过掩码时间分布蒸馏进行掩码情感表征学习。IEEE/CVF计算机视觉与模式识别会议，CVPR 2024，美国华盛顿州西雅图，2024年6月16日至22日，12830–12840。https://doi.org/10.1109/CVPR52733.2024.01219 Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J. H., & Torralba, A. (2018). 像素之声。计算机视觉—ECCV 2018—第15届欧洲会议，德国慕尼黑，2018年9月8日至14日，论文集第一部分，11205，587–604。https://doi.org/10 . 1007/978-3-030-01246-5\ 35_Zhao, K., Li, S., Cai, J., Wang, H., & Wang, J. (2019). 一种基于LSTM网络的情感符号音乐生成系统。2019年IEEE第3届信息技术与网络会议，

电子与自动化控制会议（*ITNEC*），2039–2043。https://doi.org/10.1109/ITNEC.2019.8729266Zhao, S., Ma, Y., Gu, Y., Yang, J., Xing, T., Xu, P., Hu, R., Chai, H., & Keutzer, K. (2020). 一种用于用户生成视频情绪识别的端到端视听注意力网络。*AAAI2020*，303–311。https://doi.org/10.1609/AAAI.V34I01.5364Zhou, B., Andonian, A., Oliva, A., & Torralba, A. (2018). 视频中的时间关系推理。计算机视觉—*ECCV 2018*—第*15*届欧洲会议，德国慕尼黑，*2018*年*9*月*8*日至*14*日，论文集第一部分，*11205*，831–846。https://doi.org/10.1007/978-3-030-01246-5_49Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places：一个包含 1000 万张图像、用于场景识别的数据库。*IEEE* 模式分析与机器智能汇刊，*40*(6)，1452–1464。https://doi.org/10.1109/TPAMI.2017.2723009Zhou, H., Hermans, T., Karandikar, A. V., & Rehg, J. M. (2010). 通过场景分类进行电影类型分类。*2010*年第*18*届国际多媒体会议论文集，意大利佛罗伦萨，*2010*年*10*月*25*日至*29*日，747–750。https://doi.org/10.1145/1873951.1874068Zhu, J., Sakurai, K., Togo, R., Ogawa, T., & Haseyama, M. (2024, September). Mmt-bert：一种基于多轨音乐 Transformer 和 MusicBERT 的和弦感知符号音乐生成方法。https://doi.org/10.48550/arXiv.2409.00919Zhu, X., Li, L., Liu, J., Peng, H., & Niu, X. (2018). 带有堆叠注意力模块的图像描述 Transformer。《应用科学》，*8*(5)。https://doi.org/10.3390/app8050739Zhuo, L., Wang, Z., Wang, B., Liao, Y., Bao, C., Peng, S., Han, S., Zhang, A., Fang, F., & Liu, S. (2023). 视频背景音乐生成：数据集、方法与评估。*IEEE/CVF* 国际计算机视觉会议，*ICCV 2023*，15591–15601。https://doi.org/10.1109/ICCV51070.2023.01433