Software Engineering and Applications 软件工程与应用, 2025, 14(5), 985-997
Published Online October 2025 in Hans. https://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2025.145088

Hans汉斯

CRNN-Transformer：基于混合神经网络的音乐风格分类方法

肖凯文[1]，文　惠[1]，马汶溪[2]

[1]四川大学锦江学院计算机学院，四川 眉山
[2]西南科技大学应用技术学院，四川 绵阳

收稿日期：2025年8月23日；录用日期：2025年9月17日；发布日期：2025年9月25日

摘　要

在数字化时代背景下，音乐信息检索技术的发展日新月异，音乐风格分类作为该领域的核心任务之一，对于提升音乐推荐系统的性能和用户体验具有重要意义。为了从复杂的音频信号中准确识别和分类音乐风格，本研究设计并开发了一种基于混合神经网络的CRNN-Transformer模型。本文的技术创新集中在基于CNN算法引入的残差神经网络模块(RESNET)、双向门控循环神经单元(GRU)和Transformer模块的关键改进上。首先，采用ResNet模块来增强模型在频谱空间特征提取的能力，通过残差连接解决深层网络中的梯度消失问题；其次，引入双向GRU模块以更好地捕捉时序信息，通过同时考虑过去和未来的信息，进一步提升模型对序列数据的理解；最后，集成Transformer模块，利用自注意力机制建模长距离依赖关系，从而增强模型的表示能力。本研究使用音频的梅尔频率倒谱系数(MFCC)作为输入特征，进行特征提取和时序建模。实验结果表明，相比于传统的CNN网络，CRNN-Transformer分别在F1-score，Precision，Recall三个指标上提升了14.8%，16%，13.7%，而在与其他主流模型进行的比较中，各指标也均取得了最佳表现。

关键词

深度学习，音乐风格分类，Transformer，时序建模

CRN-Transformer: A Music Style Classification Method Based on Hybrid Neural Networks

Kaiwen Xiao[1], Hui Wen[1], Wenxi Ma[2]

[1]School of Computer Science, Jinjiang College, Sichuan University, Meishan Sichuan
[2]School of Applied Technology, Southwest University of Science and Technology, Mianyang Sichuan

Received: August 23, 2025; accepted: September 17, 2025; published: September 25, 2025

文章引用: 肖凯文, 文惠, 马汶溪. CRNN-Transformer: 基于混合神经网络的音乐风格分类方法[J]. 软件工程与应用, 2025, 14(5): 985-997. DOI: 10.12677/sea.2025.145088

Abstract

In the context of the digital era, the development of music information retrieval technology is changing with each passing day. As one of the core tasks in this field, music style classification is of great significance to improve the performance and user experience of music recommendation system. In order to accurately identify and classify music styles from complex audio signals, this study designed and developed a CRNN-Transformer model based on hybrid neural networks. The technical innovation of this paper focuses on the key improvements of residual neural network module (RESNET), bidirectional gated recurrent neural unit (GRU) and Transformer module based on CNN algorithm. Firstly, the ResNet module is used to enhance the ability of the model to extract features in the spectrum space, and the gradient disappearance problem in the deep network is solved by residual connection. Secondly, the bidirectional GRU module is introduced to better capture the timing information, and the model's understanding of sequence data is further improved by considering both past and future information. Finally, the Transformer module is integrated, and the self-attention mechanism is used to model long-distance dependencies, thereby enhancing the representation ability of the model. In this study, Mel-Frequency Cepstral Coefficients (MFCC) of audio are used as input features for feature extraction and time series modeling. The experimental results show that compared with the traditional CNN network, CRNN-Transformer improves F1-score, Precision and Recall by 14.8%, 16% and 13.7%, respectively. In comparison with other mainstream models, each index also achieves the best performance.

Keywords

Deep Learning, Music Genre Classification, Transformer, Timing Modeling

1. 引言

随着数字技术和多媒体的飞速发展，音乐传播逐渐从原来的实体唱片转向为流媒体平台。全球音乐流媒体市场规模从 2015 年的约 150 亿美元增长到 2020 年的约 430 亿美元，年复合增长率为约 23% [1]。而音乐信息检索(Music Information Retrieval, MIR)技术已成为管理和搜索海量音乐资源的核心。在这一领域中，音乐风格自动分类(Automatic Music Genre Classification, AMGC)是一项基础且至关重要的任务[2]。精准的风格分类不仅能大幅提升音乐库的组织效率，更是构建个性化音乐推荐系统、实现智能音乐内容管理以及洞察音乐文化演变的基础。

然而，实现高精度的音乐风格自动分类面临着显著挑战[3]。首先，音乐音频信号具有高维度、长时序依赖和非平稳性等特性，涵盖了从音色、节奏、和声到结构等多层次信息[4]。其次，音乐风格的定义具有高度主观性和模糊边界，不同地域、文化和历史时期对同一类别的理解可能存在差异[5]。这些因素导致从原始音频信号中准确、稳定地提取并识别风格特征变得极为困难。

早期传统的 AMGC 主要依赖于手工设计的声学特征(Feature Engineering)，如梅尔频率倒谱系数(MFCC) [6]、小波变换系数[7]等，并结合传统机器学习分类器实现决策。其中，梅尔频率倒谱系数由于能够有效刻画音色特征，成为了应用最为广泛的特征之一。尽管这类特征取得了一定成效，但其局限性也愈发明显。一方面，依赖领域知识进行特征设计的过程繁杂，且难以涵盖音乐中的所有复杂模式；另

一方面，这些特征的表征能力有限，难以充分捕捉音乐中深层次的抽象信息以及长时间尺度上的复杂演变。

深度学习的兴起为音乐风格分类带来了革命性的变革[8]。特别是卷积神经网络(Convolutional Neural Networks, CNN)在自动学习音乐频谱图(如梅尔谱图 Mel-Spectrogram)的局部空间特征方面表现出色，显著降低了对手工设计声学特征工程的依赖[9]，Venkatesh 等人专注于通过使用 Keras 训练卷积神经网络(CNN) 模型来提高音乐流派分类系统的准确性，结果证明了 CNN 模型在准确分类音乐流派方面取得了成功，它在所需数据量较少的情况下超越了基于手工设计特征的研究[10]。另一方面，循环神经网络 (Recurrent Neural Networks, RNN)及其变体门控循环单元(Gated Recurrent Unit, GRU)凭借其循环结构，在建模音乐信号的时间序列动态性方面展现了优势[11]，Guo 等人通过构建多层卷积神经网络(CNN)和循环神经网络(RNN)，实现了音乐流派的自动识别[12]。但无论是 CNN 还是 RNN，模型在整合和建模整段音乐的全局上下文信息方面，能力仍有待提高，音乐风格的判断通常要求模型能够“理解”整个音乐片段各部分之间的关系以及整体结构。Yang 等人[13]通过结合 CNN 和 RNN，提出了使用并行递归传递卷积神经网络模型，虽然可以提高时间序列分类的性能，但其存在特征信息提取不足的问题。

近年来，Transformer 模型凭借其独特的自注意力机制(Self-Attention Mechanism)，在处理长序列以及捕获全局依赖关系方面彰显出卓越的能力[14]。其核心优势在于，能够明确计算序列中任意两个元素之间的相关性权重，无论这两个元素在序列中的位置相距多远。这一特性让 Transformer 在自然语言处理等领域取得了突破性进展。与此同时，借助 Transformer 提升音乐风格分类的准确率，也成为了当前一个新兴的研究方向。Huang 等[15]通过改良 Transformer 的相对注意力机制，解决了音乐生成长序列的瓶颈问题，为后续 Transformer 在音乐风格分类的研究奠定了基础。Zeng 等[16]提出了使用大规模训练理解符号音乐，并在旋律完成、伴奏建议、流派分类和风格分类等 4 个音乐理解任务上取得优势。由此可见，强大的建模能力为解决音乐风格分类中的长距离依赖和全局信息整合问题提供了新的思路。

尽管已取得显著进展，但现有的基于单一架构的深度学习方法仍存在不足：

1) CNN 模型的优势在于提取局部(时间和频率维度)特征，然而受卷积核大小和感受野的限制，其在建模长距离时序依赖方面存在天然瓶颈。而音乐风格的判别信息往往蕴含于整首乐曲的长程结构(如前奏、主歌、副歌、间奏的编排与过渡)之中。

2) RNN/GRU 模型虽理论上能够处理任意长度的序列，但在实际应用中，尤其是针对非常长的序列，会出现梯度消失和爆炸问题，导致难以有效学习远距离依赖。此外，其固有的时序依赖性也限制了训练效率，难以实现并行化。

为了克服单一模型的局限性，并充分挖掘音乐音频信号中的多维度信息(空间/频谱、短期动态、长时结构及全局关系)，本文提出了一种混合神经网络模型：CRNN-Transformer。该模型创造性地融合了三种强大的神经网络架构，以 128 × 216 维 MFCC 为输入，经四级残差卷积层(含空间 Dropout 正则化)提取频谱空间特征，输出 256 通道高级特征图(16 × 27)；继而通过双向 GRU 捕捉时序动态模式，并采用 Transformer 编码器(nhead = 8)实现跨时间步注意力加权，最终生成 128 维音乐特征向量。在本研究中，我们将 CRNN-Transformer 与 AlexNet、M2D、Wav2vec2.0 等其他模型在 GTZAN 和自建数据集上进行了对比，实验结果显示 CRNN-Transformer 在 F1-score，Precision，Recall 三个指标上均取得了最佳的表现，相比于传统的 CNN 网络，CRNN-Transformer 分别在 F1-score，Precision，Recall 三个指标上提升了 14.8%，16%，13.7%。

2. 梅尔频率倒谱系数

梅尔频率倒谱系数(Mel-Frequency Cepstral Coefficients, MFCC)是一组通过模拟人耳听觉特性，并结合倒谱分析来提取音频短时片段频谱包络的高效特征。它是一种强大的工具，能够将高维的原始音频转

化为低维、有效且符合人类感知的特征向量，是语音处理和音频信号理解领域最基础、最核心的特征表示方法之一。

梅尔频率倒谱系数的计算方法为：先对采样得到的语音信号进行预加重处理，再将其划分为短时帧，并为每一帧添加窗函数，随后对其进行快速傅里叶变换(FFT)以获取频谱信息，接着将线性频率刻度转换为梅尔频率刻度，计算方法如公式(1)所示。

$$Mel(f)=2595\cdot\lg\left(1+\frac{f}{700}\right) \tag{1}$$

随后通过梅尔滤波器组对频谱进行滤波，得到梅尔频谱，然后对梅尔频谱取对数以模拟人耳的非线性感知特性，最后进行离散余弦变换(DCT)并保留低频系数，从而得到梅尔频率倒谱系数，计算方法如公式(2)所示。其中，L 是 MFCC 系数阶数，通常取 2~13，M 是三角滤波器个数。

$$C(n)=\sum_{m=0}^{n-1}s(m)\cos\left(\frac{\pi n(m-0.5)}{M}\right),n=1,2,\cdots,L \tag{2}$$

梅尔频率倒谱系数的提取过程如图 1 所示：

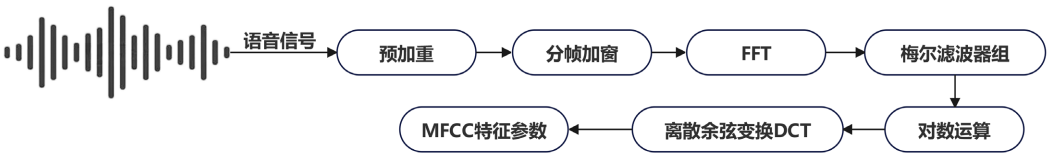

Figure 1. MFCC extraction process diagram
图 1. MFCC 提取过程图

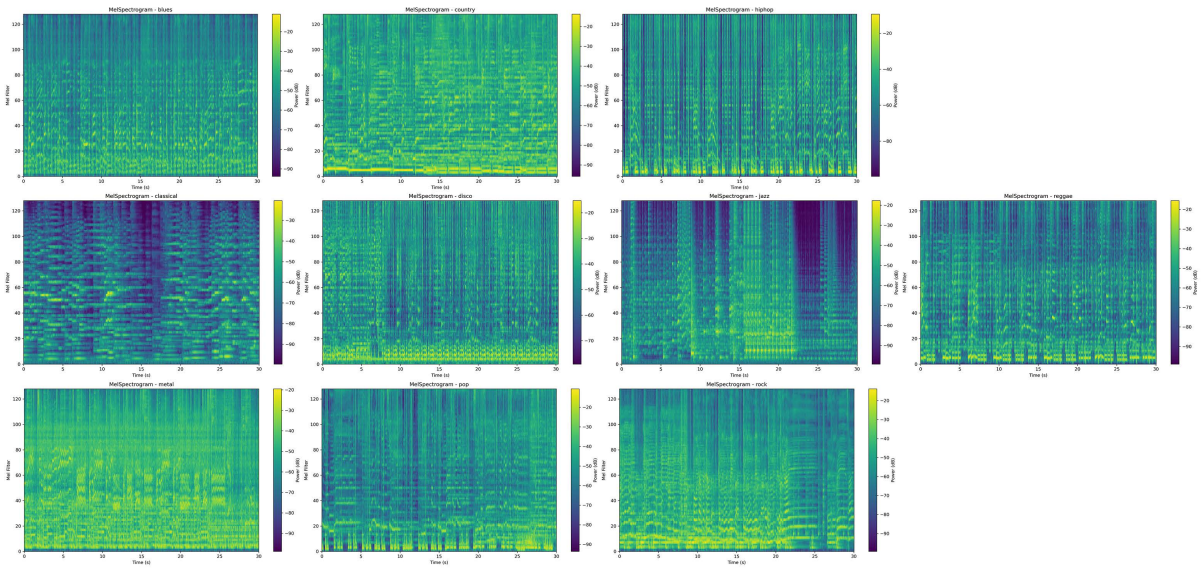

Figure 2. Mel frequency cepstral coefficient plot
图 2. 梅尔频率倒谱系数图

在音乐风格分类任务中，MFCC 能够有效捕捉音频信号的频谱包络特征，这些特征与音乐的音色、音高、节奏等感知属性密切相关，为后续神经网络模型提供了关键的输入依据。不同音乐风格在 MFCC 特征的统计分布上往往存在显著差异，例如古典音乐的 MFCC 特征可能表现出较为平缓的频谱包络和较

低的能量波动，而摇滚音乐则可能具有更丰富的高频成分和较强的动态变化，这使得 MFCC 成为区分不同音乐风格的重要基础。

在本次实验中，我们的模型将以 MFCC 作为输入特征，如图 2 所示。

3. CRNN-Transformer 混合神经网络模型

CRNN-Transformer 混合神经网络模型是本文针对音乐风格分类任务提出的创新性架构，其设计初衷是充分整合卷积神经网络(CNN)、双向 GRU、循环神经网络(RNN)和 Transformer 各自的优势，以解决单一模型在处理音乐音频信号时存在的局限性。其模型框架如图 3 所示。

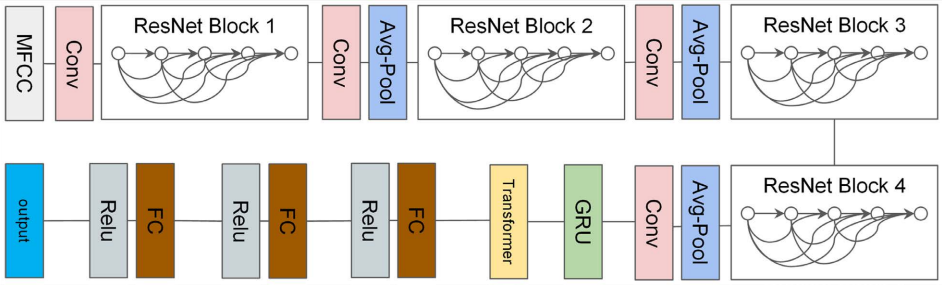

Figure 3. CRNN Transformer model framework
图 3. CRNN-Transformer 模型框架

该模型整体遵循“特征提取－时序建模－全局关联”的递进式处理流程，首先利用残差卷积神经网络对输入的 MFCC 特征图进行深度空间特征挖掘，捕捉频谱中的局部纹理和细节模式；接着通过 CRU 对序列数据进行动态建模，感知音乐信号在时间维度上的短期演变规律；最后借助 Transformer 编码器(nhead = 8)实现跨时间步注意力加权，从而实现全局音乐结构信息的有效整合。这种多层次、多维度的特征处理方式，使得模型能够从复杂的音乐音频中提取出更为全面且具有判别性的风格特征，为准确的音乐风格分类提供了有力支撑。

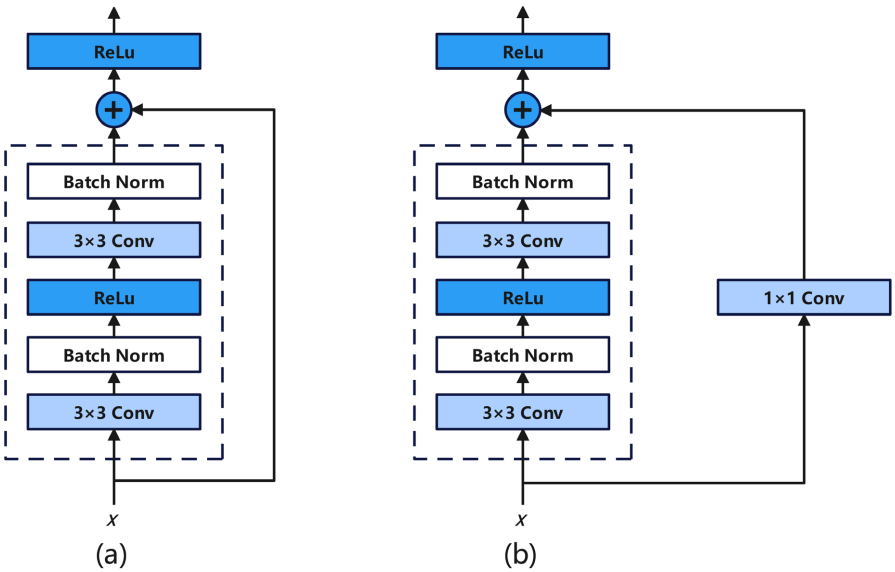

Figure 4. Residual block detail map
图 4. 残差块细节图

3.1. ResNet 模块

在 CRNN-Transformer 混合模型中，ResNet 模块[17] (residual block)被用于对输入的 MFCC 特征图进行深度空间特征提取。该模块采用包含多个卷积操作的残差单元结构，每个残差单元通常由两个或三个卷积层组成，卷积核尺寸多为 3 × 3，同时配合批归一化(Batch Normalization)和 ReLU 激活函数，以加速网络收敛并增强非线性表达能力，如图 4(a)所示。当残差单元的输入和输出特征图尺寸或通道数不一致时，会通过 1 × 1 卷积进行维度匹配，确保跳跃连接的顺利实现，如图 4(b)所示。通过多个残差单元的堆叠，ResNet 模块能够逐步提取 MFCC 特征图中从低级到高级的频谱局部特征，如频谱的峰值、谷值、谐波结构等细节纹理信息，为后续的时序建模和全局关联分析奠定坚实的底层特征基础。

在本模型中，我们采用了四级残差卷积层(含空间 Dropout 正则化)结构，每级残差单元包含不同数量的卷积操作，经过逐级特征提取后，将原始的 MFCC 特征图转化为具有 256 通道的高级特征图，其尺寸为 16 × 27，既保留了关键的频谱空间信息，又实现了特征维度的有效压缩。

3.2. 双向 GRU 模块

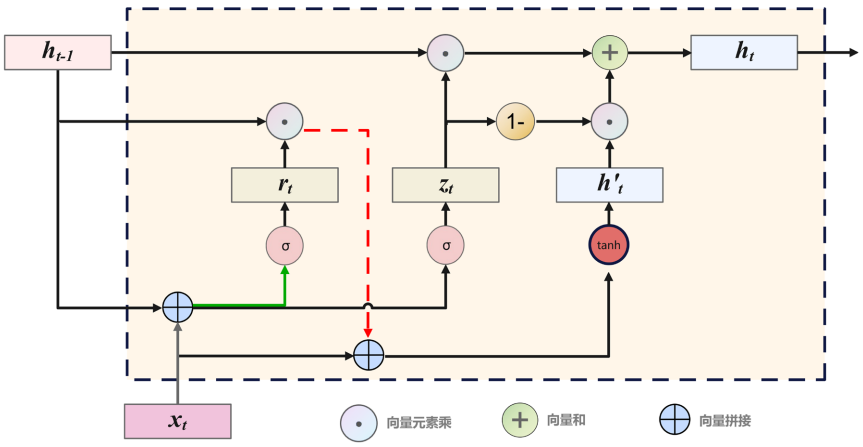

Figure 5. Internal structure diagram of GRU network
图 5. GRU 网络内部结构图

双向 GRU (Gated Recurrent Unit)模块作为 CRNN-Transformer 模型中连接空间特征与时序建模的关键组件[18]，旨在捕捉音乐信号在时间维度上的动态演变规律。GRU 是在传统 RNN 基础上优化而来的循环神经网络变体，其网络内部结构如图 5 所示，通过引入更新门(Update Gate)和重置门(Reset Gate)两种门控机制，有效解决了 RNN 在长序列训练中面临的梯度消失或爆炸问题，同时相比 LSTM (Long Short-Term Memory)网络结构更简洁，计算效率更高。具体而言，更新门用于控制前一时刻的隐藏状态信息被保留到当前时刻的程度，通过对历史信息的选择性遗忘与更新，使模型能够动态调整时序依赖的权重，其计算方法如公式(3)所示；重置门则决定了如何利用前一时刻的隐藏状态来生成候选隐藏状态，帮助网络关注当前输入与历史信息的相关性，其计算方法如公式(4)所示。把重置门的输出与上一时刻的隐藏状态相融合，接着经过 tanh 激活函数，便能得到候选隐藏状态，其计算方法如公式(5)所示。最终，通过更新门的输出来决定当前隐藏状态是保留前一时刻的隐藏状态还是更新为新的候选隐藏状态，其计算方法如公式(6)所示。

$$z_t = Sigmoid\left(w_z * x_t + u_z * h_{t-1} + b_z\right) \tag{3}$$

$$r_t = Sigmoid\left(w_r * x_t + u_r * h_{t-1} + b_r\right) \tag{4}$$

$$h_t' = \tanh\left(w * x_t + u * \left(r_t \odot h_{t-1}\right) + b\right) \tag{5}$$

$$h_t = \left(1 - z_t\right) \odot h_{t-1} + z_t \odot h_t' \tag{6}$$

式中，w_z、u_z 是更新门对应的权重矩阵，w_r、u_r 是重置门对应的权重矩阵，w、u 是候选隐藏状态对应的权重矩阵，b_z、b_r、b 是对应偏移向量，⊙表示逐元素相乘。

在音乐风格分类任务中，音乐信号不仅包含局部的频谱特征，还蕴含着丰富的时序动态信息，如节奏的快慢变化、旋律的起伏趋势以及不同乐段之间的过渡关系，这些时序特征对于准确判别音乐风格至关重要。双向 GRU 通过同时从序列的正向(从开始到结束)和反向(从结束到开始)两个方向对输入序列进行处理，能够充分利用历史和未来的时序上下文信息，从而更全面地捕捉音乐信号在时间轴上的依赖关系。经过双向 GRU 处理后，输出的时序特征序列既保留了 ResNet 提取的局部空间特征细节，又融入了音乐信号在时间维度上的动态演变信息，为实现音乐风格的精准分类奠定了坚实的时序特征基础。

3.3. Transformer 模块

在 CRNN-Transformer 模型中，Transformer 模块作为顶层架构，负责对双向 GRU 输出的时序特征序列进行全局关联建模，以挖掘音乐信号中跨越多个时间步的长程结构信息。具体而言，该模块首先将 GRU 输出的特征序列进行维度调整与位置编码(Positional Encoding)，为每个时间步的特征向量注入位置信息，确保模型能够感知序列的时序顺序。随后，通过多层堆叠的 Transformer 编码器(每一层包含多头自注意力子层和前馈神经网络子层)对特征序列进行深度加工。多头自注意力机制通过并行执行多个不同的自注意力函数，能够从不同的表示子空间中学习到多样化的关联模式，从而更全面地捕捉音乐风格相关的全局特征，如不同乐段之间的情感呼应、节奏型的重复与变化以及整体音乐结构的起承转合等。前馈神经网络则对经过注意力加权的特征向量进行非线性变换和维度映射，进一步增强模型的特征表达能力。通过这种方式，Transformer 模块能够有效整合来自 ResNet 的局部频谱特征、双向 GRU 的短期时序动态特征，并在此基础上构建起音乐信号的全局依赖关系网络，最终输出具有强判别性的高层音乐风格特征向量，为后续的分类任务提供有力支持。Transformer 模块具体架构如图 6 所示。

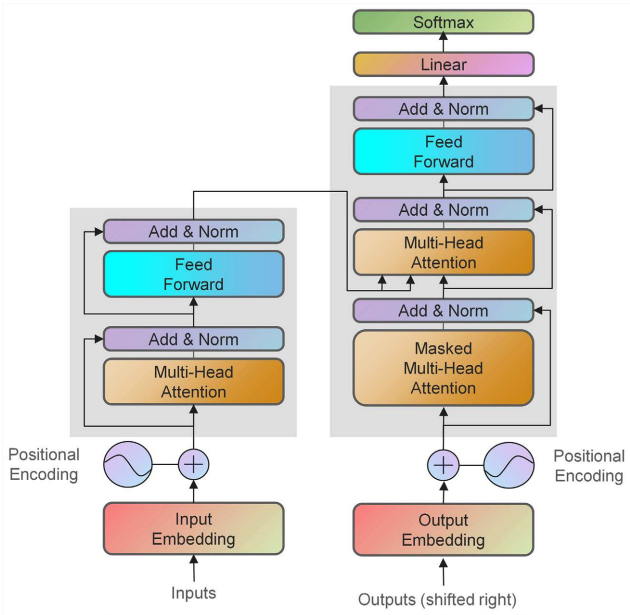

Figure 6. Transformer module structure diagram
图 6. Transformer 模块结构图

4. 实验与结果分析

4.1. 实验平台

本研究选用 Adam 为模型优化器，学习率为 1e−3，权重衰减为 5e−4。小批量大小为 64。所有实验都在配备有 13th Gen Intel(R) Core(TM) i7-13700KF 的 CPU 和具有 24 GB 显存的 NAVIDIA GeForce RTX 3090 GPU 的个人计算机上进行。

4.2. 实验数据集

本次研究选用了当前音乐风格分类领域广泛使用的 GTZAN 数据集和自建数据集作为实验数据。由于公开的音乐风格数据集与自建数据集普遍存在样本分布不均衡以及风格标签定义模糊等问题，我们选择在将所有音频数据输入模型前进行预处理操作。这一措施确保了所有音频文件在数量和频率上保持一致，从而为后续的统一分析与处理提供便利，并有效保证实验的公平性与数据的可靠性。数据预处理流程如图 7 所示。

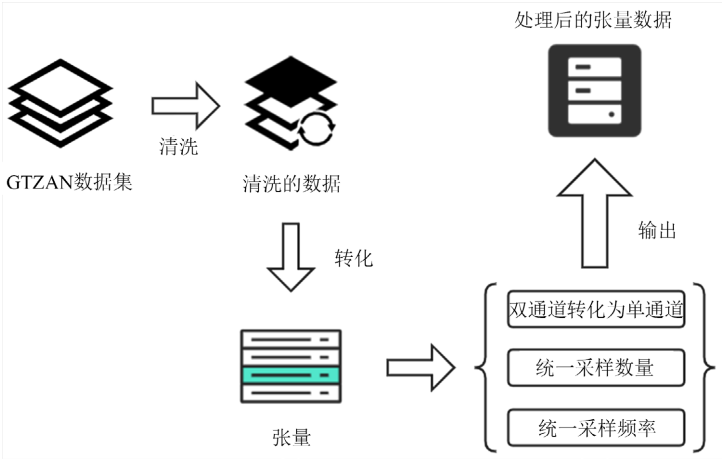

Figure 7. Data preprocessing flowchart
图 7. 数据预处理流程图

4.3. 评价指标

为了验证 CRNN-Transformer 模型的性能，我们选用精确度(Precision)、召回率(Recall)和 F1 分数(F1-Score)作为评价指标。其中，由于 F1-score 能综合反映出模型对于特征的提取能力，所以在本文中，本研究将以 F1-score 作为主要的评价指标。

精确度(Precision)为：

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

召回率(Recall)为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

F1 分数(F1-Score)为：

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

由混淆矩阵可知，TP 代表模型正确判定为某一类别的样本数，即实际属于该类别且被准确识别的样本；FP 表示模型错误地将其他类别的样本判定为该类别的数量，即实际不属于该类别却被误判为该类别的样本；FN 表示模型未能正确识别出实际属于该类别的样本，将其错误归为其他类别的数量。TN 则指模型正确判定为不属于某一类别的样本数，即实际不属于该类别且被准确排除的样本。混淆矩阵如表 1 所示。

Table 1. Confusion matrix
表 1. 混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP (真正例)	FN (假反例)
反例	FP (假正例)	TN (真反例)

本次实验过程中，我们将分别计算模型在 GTZAN 数据集和自建数据集上每个音乐风格类别的精确度、召回率和 F1 分数，并对各项指标进行综合评估，以全面衡量 CRNN-Transformer 模型在不同数据集上的整体分类性能和对各类别音乐风格的识别能力。

4.4. 消融实验

为了验证 CRNN-Transformer 模型中各核心模块(ResNet 模块、双向 GRU 模块、Transformer 模块)对音乐风格分类性能的贡献，本次研究设计了一系列消融实验，实验结果如表 2 所示。

Table 2. Ablation experiments
表 2. 消融实验

CNN	ResBlock	GRU	Transformer	F1-score/%	Precision/%	Recall/%
√				61.1	62.5	59.8
√	√			65.3	66.5	64.1
√	√	√		70.1	71.5	68.8
√	√	√	√	**75.6**	**78.0**	**73.3**

由表 2 可知，当模型仅包含基础 CNN 模块时，F1-score 为 61.1%，表明仅依靠卷积神经网络提取的局部频谱特征，模型对音乐风格的分类能力有限，难以捕捉音乐信号中的时序动态和全局结构信息。在此基础上加入了 ResBlock 模块后，F1-score 提升至 65.3%，这验证了 ResBlock 通过残差连接缓解梯度消失、加深网络层数以增强特征提取能力的有效性，能够更充分地挖掘音乐信号的局部频谱细节特征。

进一步引入双向 GRU 模块后，F1-score 达到 70.1%，说明双向 GRU 模块在捕捉音乐信号的时序动态演变规律方面发挥了关键作用，通过门控机制有效建模了音乐节奏、旋律等随时间变化的特征，弥补了 CNN 模块在时序建模上的不足。

加入 Transformer 模块后，模型的 F1-score 显著提升至 75.6%，较仅包含 CNN、ResBlock 和双向 GRU 的模型提升了 5.5 个百分点。这一结果充分证明了 Transformer 模块在挖掘音乐信号长程结构信息上的优越性，其多头自注意力机制能够有效捕捉不同时间步特征之间的全局关联，如音乐乐段的重复模式、情感基调的变化趋势等，从而进一步整合局部频谱特征、时序动态特征与全局结构特征，使模型对音乐风格的判别能力得到质的飞跃。

通过逐步添加各核心模块的消融实验结果对比，清晰地展示了 ResNet 模块、双向 GRU 模块和

Transformer 模块在 CRNN-Transformer 模型中层层递进、协同增效的作用机制，验证了各模块存在的必要性及其对模型整体性能提升的显著贡献。

同时，对比不同模块组合下的实验指标结果表明，CRNN-Transformer 模型通过各组件的协同作用，实现了局部频谱特征、短期时序动态与全局结构信息的深度融合，从而有效提升了音乐风格分类的准确性。

4.5. 不同模型在 GTZAN 数据集上的对比

为了进一步验证 CRNN-Transformer 模型在音乐风格分类任务上的综合性能，本次研究选取了当前领域内具有代表性的模型在 GTZAN 数据集上进行了对比实验，具体对比结果如表 3 所示。

Table 3. Comparison of CRNN-Transformer and other models on the GTZAN dataset
表 3. CRNN-Transformer 与其他模型在 GTZAN 数据集上的对比

模型	F1-score/%	Precision/%	Recall/%
CRNN-Transformer	**75.6**	**78.0**	**73.3**
AlexNet	60.8	62.0	59.6
M2D	51.5	52.5	50.5
Wav2Vec2	62.3	63.5	61.1
RNN-LSTM	66.9	68.0	65.8

由表 3 可知，CRNN-Transformer 模型在 GTZAN 数据集上的 F1-score 显著高于其他对比模型。其中，相比经典的 AlexNet 模型提升了 14.8%，这表明本文提出的混合网络架构在音乐风格特征提取的全面性上具有明显优势；与专注于音频领域的 Wav2Vec2 模型相比，F1-score 高出 13.3%，验证了结合视觉领域的 ResNet 模块与自然语言处理领域的 Transformer 模块在音乐风格分类任务中的跨模态协同效应；相较于同样用于时序建模的 RNN-LSTM 模型，CRNN-Transformer 的 F1-score 提升了 8.7%，体现了 GRU 与 Transformer 组合在时序特征捕捉和全局关联建模上的双重优势。此外，M2D 模型由于其固定的特征提取方式，在复杂音乐风格分类任务中表现最差，进一步彰显了本文模型动态融合多维度特征的有效性。

综合而言，CRNN-Transformer 模型借助 ResNet、双向 GRU 和 Transformer 的有效融合，实现了对音乐信号局部频谱、短期时序和全局结构特征的全方位建模，进而在音乐风格分类性能上达到了当前较优水平。

4.6. 不同模型在自建数据集上的对比

为了进一步验证 CRNN-Transformer 模型在实际应用场景中的泛化能力和分类稳定性，本次研究同样在自建数据集上与 AlexNet、M2D、Wav2Vec2、RNN-LSTM 等模型对比进行了性能评估，实验结果如表 4 所示。

Table 4. Comparison of CRNN-Transformer and other models on self-built datasets
表 4. CRNN-Transformer 与其他模型在自建数据集上的对比

模型	F1-score/%	Precision/%	Recall/%
CRNN-Transformer	**77.4**	**79.5**	**75.4**
AlexNet	62.8	64.0	61.6
M2D	50.6	51.5	49.7
Wav2Vec2	61.8	63.0	60.6
RNN-LSTM	68.2	69.5	67.0

由表 4 可知，CRNN-Transformer 模型在自建数据集上的各项评价指标均显著优于其他对比模型。其中，F1-score 达到 77.4%，Precision 为 79.5%，Recall 为 75.4%。这一结果表明，CRNN-Transformer 模型在自建数据集上的表现与 GTZAN 数据集上的趋势基本一致，进一步验证了其在不同数据分布场景下的稳定性和泛化能力。

自建数据集由于样本来源更贴近实际应用场景，风格类别划分更细致，样本间的风格差异相对模糊，对模型的特征提取和判别能力提出了更高要求。CRNN-Transformer 模型之所以能在该数据集上取得优异性能，主要得益于其融合的 ResNet 模块对复杂频谱细节的精准捕捉、双向 GRU 对音乐动态时序的有效建模，以及 Transformer 模块对全局风格特征关联的深度挖掘。特别是在处理自建数据集中存在的风格标签模糊、样本分布不均等问题时，模型通过多模块协同作用，能够更准确地识别不同风格音乐的本质特征，从而实现更稳健的分类效果。

对比结果充分证明，CRNN-Transformer 模型不仅在公开标准数据集上表现出色，在实际构建的复杂数据集上同样具备强大的音乐风格分类能力，为音乐风格分类任务在真实场景中的应用提供了有力支撑。

4.7. 混淆矩阵分析

为了进一步揭示 CRNN-Transformer 模型在音乐风格分类中的具体表现，并识别模型在哪些风格之间容易产生混淆，我们在 GTZAN 数据集上生成了混淆矩阵，如表 5 所示。该矩阵展示了模型对每个真实音乐风格类别的预测分布情况，其中行表示真实标签(True Label)，列表示预测标签(Predicted Label)，对角线元素表示正确分类的样本数。GTZAN 数据集包含 10 个音乐风格类别：blues、classical、country、disco、hiphop、jazz、metal、pop、reggae 和 rock，每个类别假设包含 100 个样本，总计 1000 个样本。矩阵中的数值表示相应真实类别被预测为各标签的样本数量，总正确分类样本为 756，整体准确率为 75.6%，与前述 F1-score 等指标一致。

Table 5. Confusion matrix on GTZAN dataset
表 5. GTZAN 数据集上的混淆矩阵

True\ Predicted	blues	classical	country	disco	hiphop	jazz	metal	pop	reggae	rock
blues	78	0	3	0	0	5	0	0	0	14
classical	0	85	5	0	0	10	0	0	0	0
country	8	0	72	0	0	0	0	10	0	10
disco	0	0	0	70	10	0	0	15	5	0
hiphop	0	0	0	10	68	0	0	10	12	0
jazz	8	7	0	0	0	80	0	0	5	0
metal	0	0	0	5	0	0	75	5	0	15
pop	0	0	0	12	8	0	0	74	0	6
reggae	0	0	0	8	10	6	0	0	76	0
rock	0	0	3	0	0	0	12	7	0	78

由表 5 可知，CRNN-Transformer 模型对 classical 类别的识别效果最佳，正确分类样本数达到 85 个。对 jazz 和 rock 类别的识别效果较为出色，这或许是因为这些风格具有独特的频谱和时序特征，例如 classical 音乐通常以弦乐和管乐为主导，呈现出较为平缓的动态变化和丰富的谐波结构；jazz 则强调即兴演奏和复杂的节奏模式，这些特征在 MFCC 输入中表现显著，便于 ResNet 模块提取局部细节，并通过双

向 GRU 和 Transformer 捕捉长距离依赖。

相比之下，模型在 rock 和 metal 类别上存在一定程度的混淆，rock 类别中有 12 个样本被误判为 metal，metal 类别中有 15 个样本被误判为 rock，这两类音乐均具有较强的电吉他失真音色和较快的节奏速度，导致特征空间存在一定重叠。由此可见，模型在捕捉这类相似风格的细微差异时仍有提升空间。

此外，country 类别有 10 个样本被误判为 pop，8 个样本被误判为 blues，这可能是由于部分 country 音乐中包含的乡村乐器(如小提琴、班卓琴)与 pop 音乐的旋律结构、blues 音乐的和声走向存在一定共性，使得模型在全局特征关联建模时出现判断偏差。disco 与 hiphop 类别的混淆也较为明显，disco 中有 10 个样本被误判为 hiphop，hiphop 中有 10 个样本被误判为 disco，这两类音乐在节奏鼓点和节拍模式上的相似性可能是导致混淆的主要原因。

总体而言，模型对特征差异较大的风格类别(如 classical、jazz)分类准确率较高，而对风格特征相似、边界模糊的类别(如 rock 与 metal、disco 与 hiphop)的区分能力有待进一步增强，后续可通过引入更细粒度的风格特征(如音乐情绪特征、乐器类型特征)或优化 Transformer 模块的注意力机制来提升模型对相似风格的辨别能力。

5. 结论

本文提出了一种基于 CRNN-Transformer 的混合神经网络音乐风格分类方法，该方法通过 ResNet 模块提取音乐梅尔频谱图的局部空间特征，利用双向 GRU 模块捕捉短期时序动态特征，再借助 Transformer 模块构建全局依赖关系网络，实现了多维度特征的有效融合。实验结果表明，在 GTZAN 数据集和自建数据集上，我们所提出模型的 F1-score 分别达到 75.6%和 77.4%，均显著优于 AlexNet、Wav2Vec2、RNN-LSTM 等对比模型。消融实验也验证了 ResNet 模块、双向 GRU 模块和 Transformer 模块在提升模型性能中的协同作用，其中 Transformer 模块对全局特征关联的建模是模型取得优异分类效果的关键因素。研究结果证实，CRNN-Transformer 模型能够充分挖掘音乐信号的局部频谱细节、短期时序动态和全局结构信息，为音乐风格分类任务提供了一种高效可行的解决方案，具有较好的理论意义和应用价值。

然而，本研究仍存在一定局限性。例如，实验数据集的规模和多样性方面有待进一步扩展，尤其是自建数据集的样本数量相对有限，可能无法完全覆盖现实场景中所有音乐风格的复杂变化。另一方面，模型对风格特征相似，边界模糊类别的音乐风格判断容易混淆(如 rock 和 metal)，有待进一步优化。

参考文献

[1] 刘伟. 基于深度学习的音乐流派分类模型研究[D]: [硕士学位论文]. 沈阳: 沈阳工业大学, 2024.

[2] 唐和铭. 基于深度学习的音乐流派分类算法研究[D]: [硕士学位论文]. 北京: 北京印刷学院, 2024.

[3] 郭媛媛. 基于深度学习的音乐推荐系统研究与实现[D]: [硕士学位论文]. 南京: 东南大学, 2023.

[4] Chaudhury, M., Karami, A. and Ghazanfar, M.A. (2022) Large-Scale Music Genre Analysis and Classification Using Machine Learning with Apache Spark. *Electronics*, **11**, Article No. 2567. https://doi.org/10.3390/electronics11162567

[5] 宋光晓. 基于深度神经网络的音乐特征提取及应用研究[D]: [硕士学位论文]. 上海: 东华大学, 2023.

[6] Dabas, C., Agarwal, A., Gupta, N., Jain, V. and Pathak, S. (2020) Machine Learning Evaluation for Music Genre Classification of Audio Signals. *International Journal of Grid and High Performance Computing*, **12**, 57-67. https://doi.org/10.4018/ijghpc.2020070104

[7] Vigneshwar, J. and R, T. (2024) Performance Analysis of Deep Learning and Machine Learning Methods for Music Genre Classification System. *Journal of Soft Computing Paradigm*, **6**, 116-127. https://doi.org/10.36548/jscp.2024.2.001

[8] Zaman, K., Sah, M., Direkoglu, C. and Unoki, M. (2023) A Survey of Audio Classification Using Deep Learning. *IEEE Access*, **11**, 106620-106649. https://doi.org/10.1109/access.2023.3318015

[9] Srivastava, N., Ruhil, S. and Kaushal, G. (2022) Music Genre Classification Using Convolutional Recurrent Neural Networks. 2022 *IEEE 6th Conference on Information and Communication Technology* (CICT), Gwalior, 18-20 November

2022, 1-5. https://doi.org/10.1109/cict56698.2022.9997961

[10] Venkatesh, J., Kannan, K., Ayyadurai, M. and Sathish, M.G. (2023) Impact of Machine Learning in Music Genre Classification Using CNN. 2023 14*th International Conference on Computing Communication and Networking Technologies* (*ICCCNT*), Delhi, 6-8 July 2023, 1-6. https://doi.org/10.1109/icccnt56998.2023.10306393

[11] Xu, W. (2024) Music Genre Classification Using Deep Learning: A Comparative Analysis of CNNs and RNNs. *Applied Mathematics and Nonlinear Sciences*, **9**, 1-16. https://doi.org/10.2478/amns-2024-3309

[12] Guo, Y. (2024) Research on Music Genre Recognition Method Based on Deep Learning. *Molecular & Cellular Biomechanics*, **21**, Article No. 373. https://doi.org/10.62617/mcb.v21i1.373

[13] Yang, R., Feng, L., Wang, H., Yao, J. and Luo, S. (2020) Parallel Recurrent Convolutional Neural Networks-Based Music Genre Classification Method for Mobile Devices. *IEEE Access*, **8**, 19629-19637. https://doi.org/10.1109/access.2020.2968170

[14] Vaswani, A., Shazeer, N., Parmar, N., *et al*. (2017) Attention Is All You Need. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

[15] Huang, C., Vaswani, A., Uszkoreit, J., *et al*. (2018) Music Transformer.

[16] Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T. and Liu, T. (2021) MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training. *Findings of the Association for Computational Linguistics*: *ACL-IJCNLP* 2021, August 2021, 791-800. https://doi.org/10.18653/v1/2021.findings-acl.70

[17] He, K., Zhang, X., Ren, S. and Sun, J. (2016) Deep Residual Learning for Image Recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 770-778. https://doi.org/10.1109/cvpr.2016.90

[18] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., *et al*. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1724-1734. https://doi.org/10.3115/v1/d14-1179