基于深浅卷积注意力的改进 Transformer 的音乐流派分类研究

彭 扬[1] 李昌翰[1*] 李旭航[1] 叶星吟[1] 颜镜竹[1] 杨羽西[1] 范晨希[1]
PENG Yang LI Changhan LI Xuhang YE Xingyin YAN Jingzhu YANG Yuxi FAN Chenxi

摘 要 随着数字音乐的快速发展，海量的音乐作品不断涌现，如何高效、准确地对音乐流派进行自动分类成为智能音乐推荐、检索和版权管理中的关键问题。传统基于手工特征或浅层神经网络的方法在复杂、多样的音乐数据集上往往难以获得理想性能。针对这一问题，文章提出一种基于深浅卷积注意力的改进 Transformer 与 ArcFace 的音乐流派分类方法。此方法首先利用深、浅两级 ResNet 提取多尺度特征，并通过多头注意力机制增强时序特征建模能力，从而获得更具判别性的全局特征表示。随后引入 ArcFace 分类器提升类间可分性，并结合交叉熵损失与三元组损失进行联合优化。实验结果表明，所提方法相较于传统的机器学习与深度学习模型具有更高的分类性能，验证了所提方法在音乐流派分类任务中的有效性和鲁棒性。

关键词 Transformer；ResNet；多头注意力；ArcFace

doi：10.3969/j.issn.1672-9528.2026.06.010

0 引言

在数字音乐快速发展的背景下，音乐流媒体平台的普及带来了前所未有的音频资源获取方式，但也随之产生了对自动化管理和智能分析的迫切需求。尤其在音乐信息检索（music information retrieval, MIR）领域，如何准确识别和分类音乐流派已成为一项关键任务。这不仅关系到个性化推荐、版权保护与内容管理等应用场景，也反映了深度学习在非结构化音频数据处理上的能力与挑战。传统依赖人工设计特征或浅层模型的方法在处理复杂多样的音乐信号时往往表现不足，难以兼顾分类的准确性与泛化性[1]。因此，探索能够更好建模音频特征的深度学习方法，已成为提升音乐流派自动分类性能的重要方向[2]。

近年来，研究者针对音乐流派分类的多样化挑战提出了多种深度学习方法以提升分类性能。Dong[3] 针对传统方法难以捕捉时序依赖的问题，将整段音频划分为 3 s 片段，并利用卷积神经网络（CNN）学习局部时频特征后通过投票实现

1. 重庆科技大学 重庆 401331
[基金项目] 重庆市社科联博士项目（2023BS035）；重庆市教委科学技术研究项目（KJQN202301503）

参考文献：

[1] 唐箭 . 提示工程赋能老年教育治理研究 [J]. 科技资讯 , 2025, 23(16):242-245.

[2] 唐箭 . 自然语言处理技术在老年教育治理中的应用研究 [J]. 科技资讯 , 2025,23(15):218-221.

[3] 王洋 , 李代萍 , 杨燕武 , 等 . 人工智能在老年医学教学中的应用现状及前景 [J]. 实用老年医学 ,2025,39(7):753-756.

[4] 唐箭 . 基于腾讯混元大模型的老年教育智能体构建与应用研究 [J]. 科技资讯 ,2025,23(13):224-229.

[5] 张国丰 , 马玉环 . AI 背景下老年思政教育创新路径研究 [J]. 吉林广播电视大学学报 ,2025(3):73-75.

[6] 景燕敏 . 基于智能化技术的老年教育有声在线教学系统设计 [J]. 软件 ,2024,45(7):32-34.

[7] 赵洋洋 . 人工智能环境下老年群体协作学习环境建设路径 [J]. 中国成人教育 ,2022(24):29-33.

[8] 王思瑶 , 马秀峰 . 人工智能与老年教育深度融合研究 [J]. 成人教育 ,2022,42(9):43-50.

【作者简介】

祁萌（1997—），女，黑龙江绥化人，硕士研究生，助理实验师，研究方向：老年教育、实验室管理。

（收稿日期：2025-10-27 修回日期：2026-05-29）

整曲分类，使得 GTZAN 数据集上的准确率达到了接近人类水平的 70%。为解决 CNN 感受野有限、难以捕捉全局依赖的问题，Zhuang 等 [4] 引入 Transformer 架构，通过自注意力机制对谱图长程依赖进行建模，实验结果显示在 GTZAN 上获得 76.1% 的准确率。在时域特征学习方面，Allamy 等 [5] 设计了基于残差块的一维卷积神经网络（1D-CNN），能够直接从原始波形中提取判别性特征，减少了频谱转换和人工特征提取环节，在 GTZAN 上达到 80.93% 的平均准确率。针对开放集识别问题，Park 等 [6] 基于 EfficientNet-B3 结合 PROSER 算法提出方案，既能区分已知流派也能识别未知流派，整体准确率为 74.1%。为提升模型的全局特征提取能力，Wu 等 [7] 将 Vision Transformer 与卷积及多通道注意力机制结合，从而增强局部和全局特征交互学习，在 GTZAN 上实现 86.8% 的准确率。此外，Li 等 [8] 提出 CNN+MLP 的双分支输入架构，利用 Mel 频谱和统计特征的多模态融合来缓解单一特征空间的欠拟合，在 GTZAN 上取得 77% 的准确率并验证了该策略的有效性。

尽管已有方法在分类准确率上取得了提升，但现有研究仍面临一些挑战。例如数据集规模较小，难以全面覆盖多样化的音乐类型，深度学习模型对少数类样本敏感等问题。针对上述问题，本文提出一种结合深浅卷积特征与改进 Transformer 的方法，旨在同时增强局部特征提取与全局依赖建模能力，并通过判别性损失优化提升类间区分度，从而改善多流派复杂场景下的分类性能。

1 相关技术

1.1 ResNet 特征提取网络

残差网络通过引入恒等映射的残差连接，有效缓解了深层网络训练的梯度消失问题，使得网络能够训练得更深 [9]。模型由两层卷积和非线性激活组成，并通过跨层恒等映射将输入 $\boldsymbol{x}$ 与卷积块输出相加，残差函数用公式表示为：

$$\boldsymbol{y}=F\left(\boldsymbol{x},\{\boldsymbol{W}_i\}\right)+\boldsymbol{x} \tag{1}$$

式中：$F\left(\boldsymbol{x},\{\boldsymbol{W}_i\}\right)$为残差分支上的卷积操作，$\boldsymbol{x}$为输入，$\boldsymbol{y}$为输出。

本文采用深、浅两级 ResNet 以提取多尺度特征，深层 ResNet 输出语义特征用于生成查询向量，浅层 ResNet 输出局部特征用于生成键和值。

1.2 Transformer 与多头自注意力机制

Transformer 是一种基于自注意力机制的序列建模架构，由编码器和解码器组成 [10]。其中每个编码器层包含多头自注意力模块、前馈网络（feed-forward network, FFN）、残差连接和层归一化（LayerNorm）。其单头注意力计算公式为：

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{2}$$

式中：$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 分别为查询、键和值，d_k 为键的维度。

多头注意力通过并行计算多个头的注意力后拼接得到最终表示为：

$$\text{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\text{head}_1,\ldots,\text{head}_h\right)\boldsymbol{W}^O \tag{3}$$

由此可以让模型从不同子空间捕捉多样化的依赖关系。标准 Transformer 架构已在自然语言处理、语音识别和音频分类任务中取得良好效果，但直接用于谱图建模仍存在局限，如对局部纹理的捕捉不足等问题，这正是本文在后续章节提出改进的动机。

2 模型设计

2.1 多尺度特征提取与改进 Transformer

输入音频首先经短时傅里叶变换转化为 Mel 频谱图，形成 $T\times F$ 的二维时频矩阵，作为后续网络的输入。多尺度特征提取模块由两条不同深度的 ResNet 分支组成。深层分支包含多个残差模块，提取高层语义特征，能够表达整体音色和曲式信息，浅层分支仅包含较少卷积层，保留更多局部细节，如短时能量变化和瞬态纹理。两路特征经拼接后通过线性变换统一维度，得到适合 Transformer 处理的特征序列。

改进的 Transformer 编码模块在输入端加入特征融合层，使深浅卷积提取的多尺度特征能够在进入注意力计算前充分融合。随后，特征序列经过多头自注意力、残差连接、层归一化和前馈网络等模块处理，通过堆叠多层编码器实现全局依赖建模。多头注意力可以从不同子空间学习特征相关性，提升模型捕捉复杂音乐模式的能力，残差连接和层归一化保证训练过程稳定，避免梯度消失和收敛缓慢。与标准 Transformer 相比，本文模型能够更好地整合局部和全局信息，提升谱图特征的判别性。

2.2 判别性分类优化模块

为了进一步提高分类器的区分能力，本文在编码器输出后引入判别性分类优化模块。首先，采用 ArcFace 分类器对嵌入特征进行角度间隔约束，使类内样本更加紧凑、类间边界更加清晰，从而提升最终的分类性能。ArcFace 的目标函数用公式表示为：

$$L_{\text{ArcFace}}=-\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp\left(s\cos\left(\theta_{y_i}+m\right)\right)}{\exp\left(s\cos\left(\theta_{y_i}+m\right)\right)+\sum_{j\neq y_i}\exp\left(s\cos\theta_j\right)}\right) \tag{4}$$

式中：s 为缩放因子，m 为角度间隔超参数，θ_{y_i}为第 j 个样本与其真实类别权重向量的夹角。该损失通过在角度空间引入间隔约束，能够显著增强特征嵌入的判别性。

此外，为了进一步优化特征空间分布，本文在训练过程中引入三元组损失，通过约束锚点、正样本和负样本的相对

距离，使同类样本之间更加接近，异类样本之间更加远离：

$$L_{\text{triplet}} = \sum_{i=1}^{N} \text{Big}\left[\left\| f\left(x_i^a\right) - f\left(x_i^p\right) \right\|_2^2 - \left\| f\left(x_i^a\right) - f\left(x_i^n\right) \right\|_2^2 + \alpha \text{Big}\right]_+ \tag{5}$$

最终，联合优化目标为交叉熵损失、ArcFace 损失和三元组损失的加权和为：

$$L_{\text{total}} = \lambda_1 L_{\text{CE}} + \lambda_2 L_{\text{ArcFace}} + \lambda_3 L_{\text{triplet}} \tag{6}$$

通过该优化策略，模型不仅能获得较高的分类准确率，还能在特征空间形成良好的类间分布，增强泛化能力。模型具体流程如图 1 所示。

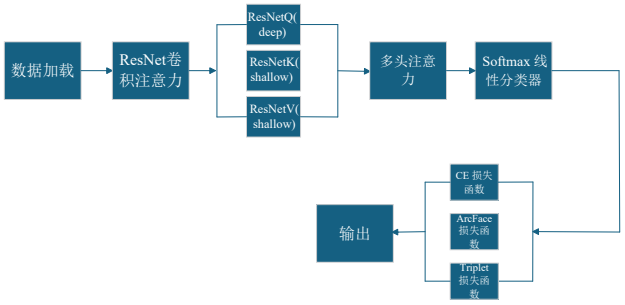

图 1 模型流程图

3 实验设计与结果分析

3.1 实验数据

本文选用的实验数据集为 GTZAN，该数据集是音乐信息检索领域的经典基准，包含 1 000 首长度为 30 s 的音乐片段，分为 10 个流派类别：Blues、Classical、Country、Disco、Hiphop、Jazz、Metal、Pop、Reggae、Rock，每类各 100 首。为避免数据泄漏，本文严格按照 7:1:2 的比例划分训练集、验证集和测试集。

3.2 实验对比

为验证所提改进 Transformer 模型的有效性，本文选取多种主流深度学习方法进行对比实验。对比方法包括：1D-CNN 模型 [5]、Improved ViT 模型 [1]、基于 Transformer 的分类方法 [4]、EfficientNet-B3 模型 [6]。所有对比模型均在相同的 GTZAN 数据集上进行训练和测试，输入特征为 Mel 频谱图，数据预处理、训练参数和优化策略保持一致，以保证对比的公平性。

3.3 评价指标

在多分类任务中，评价指标源自混淆矩阵，可从不同角度反映模型性能。本文采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）、F_1 值（F_1-score）与 AUC 进行综合评估，其中除准确率外，其余指标均采用宏平均（macro-average）。准确率用于衡量总体预测正确的比例，即所有被正确分类的样本数占总样本数的比值：

$$\text{ACC} = \frac{\sum_{c=1}^{C} \text{TP}_c}{N} \tag{7}$$

精确率是度量正类预测的可靠性，即被判为第 c 类的样本中有多少是真正属于该类，多分类取宏平均：

$$\text{Precision} = \frac{1}{C}\sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} \tag{8}$$

召回率是度量对正类样本的覆盖能力，即真实属于第 c 类的样本中被正确识别的比例，多分类取宏平均：

$$\text{Recall} = \frac{1}{C}\sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} \tag{9}$$

F_1 值是精确率与召回率的调和平均，综合反映预测纯度与覆盖率，多分类取宏平均：

$$F_1 = \frac{1}{C}\sum_{c=1}^{C} \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{10}$$

AUC 最终以 ROC 曲线衡量整体区分能力，多分类采用“一对其余”方式得到各类 AUC，再作宏平均：

$$\text{AUC}_c = \int_0^1 \text{TPR}_c\left(\text{FPR}\right)\text{dFPR} \tag{11}$$

其中，TPR 与 FPR 分别为：

$$\text{TPR} = \frac{\text{TP}}{\text{TP+FN}},\quad \text{FPR} = \frac{\text{FP}}{\text{FP+TN}} \tag{12}$$

3.4 实验结果

为验证本文所提改进 Transformer 模型的性能优势，将其与多种对比方法在相同实验条件下进行测试，结果如表 1 所示。

表 1 实验结果

模型	Accuracy	Precision	Recall	F_1-score	AUC
1D-CNN	0.809 3	0.815 4	0.820 0	0.821 5	0.941 0
Improved ViT	0.868 0	0.851 8	0.855 0	0.854 3	0.958 0
Transformer	0.761 0	0.775 6	0.778 0	0.776 9	0.973 2
EfficientNet-B3	0.741 0	0.788 1	0.791 0	0.790 8	0.981 0
本文方法	0.905 0	0.903 7	0.905 0	0.905 4	0.987 3

从表 1 可以看出，各模型在 Accuracy、Precision、Recall、F_1-score 和 AUC 五个指标上均取得了较为理想的结果。其中，卷积神经网络（1D-CNN）和 EfficientNet-B3 在整体性能上略低于基于注意力机制的方法，表明注意力机制在捕捉全局依赖方面对分类任务具有优势。相比之下，Improved ViT 在各项指标上明显优于标准 Transformer，说明模型结构的改进能够有效提升分类性能。总体而言，本文提出的方法在所有指标上均取得了最佳表现，验证了所提方法的有效性和优越性。

为了进一步展示分类结果，绘制了改进 Transformer 模型

在测试集上的混淆矩阵，如图 2 所示，其中，坐标中的 0~9 分别代表 Blues、Classical、Country、Disco、Hiphop、Jazz、Metal、Pop、Reggae、Rock 十种不同类型的音乐风格。混淆矩阵显示各流派样本主要分布在对角线位置，不同流派之间存在一定混淆，其中 Disco，Reggae 和 Rock 的混淆较为明显，说明这些类别在特征空间存在相似性。

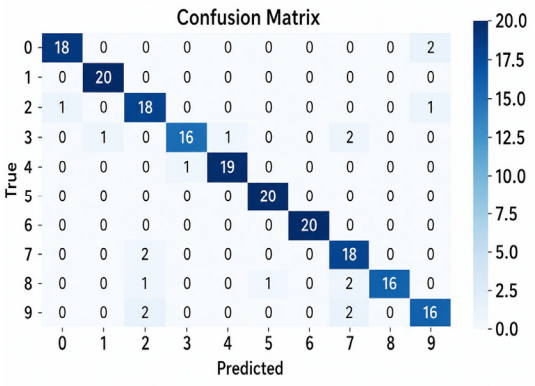

图 2 混淆矩阵

4 结论与展望

本文围绕音乐流派分类任务，提出了一种基于多尺度卷积特征和改进 Transformer 编码器的分类方法。首先，通过深浅双路卷积网络提取多尺度谱图特征，在保留局部纹理信息的同时增强高层语义表达。随后，引入改进的 Transformer 编码器对多尺度特征进行全局依赖建模；最后，结合角度间隔约束的分类器和三元组损失优化特征分布，提高类别间可分性。在 GTZAN 数据集上的实验表明，该方法在 Accuracy、Precision、Recall、F_1-score 和 AUC 等指标上均取得较高数值。通过混淆矩阵分析，不同流派的样本主要集中在对角线附近，部分流派之间存在混淆，说明在时频特征相似度较高的类别之间仍存在判别难度。后续研究可探索更高效的特征增强和时序建模策略，以提升相似流派之间的区分能力。或者考虑引入更大规模的音乐数据集，增强模型的泛化性能等进一步拓展应用价值。

参考文献：

[1]Zaman K, Sah M, Direkoglu C, et al. A survey of audio classification using deep learning[J]. IEEE Access, 2023, 11: 106620-106649.

[2]LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[3]]Dong Mingwen. Convolutional neural network achieves human-level accuracy in music genre classification[PP/OL]. (2018-02-27)[2026-05-23].https://doi.org/10.48550/arXiv.1802.09697.

[4]Zhuang Yingying, Chen Yuezhang, Zheng Jie. Music genre classification with transformer classifier[C]//ICDSP'20: Proceedings of the 2020 4th International Conference on Digital Signal Processing. NewYork: ACM, 2020: 155-159.

[5]Allamy S, Koerich A L. 1D CNN architectures for music genre classification[C/OL]//2021 IEEE symposium series on computational intelligence(SSCI).Piscataway:IEEE, 2021[2026-05-23].https://www.bohrium.com/paper-details/1d-cnn-architectures-for-music-genre-classification/867763329311965202-108628.

[6]Park K, Jeon J, Park S, et al. Music classification scheme based on EfficientNet-B3[J/OL]. Human-centric Computing and Information Sciences, 2023[2026-05-23].https://pure.dongguk.edu/en/publications/music-classification-scheme-based-on-efficientnet-b3/.

[7]Wu Pingping, Gao Weijie, Chen Yitao, et al. An improved ViT model for music genre classification based on mel spectrogram[J]. PloS One, 2025, 20(3): e0319027.

[8]Li Xiang, Li Fan, Lu Zhipeng, et al. Music genre classification: a comprehensive study on feature fusion with CNN and MLP architectures[J]. Applied and Computational Engineering, 2025, 132(1): 159-166.

[9]He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Piscataway: IEEE, 2016: 770-778.

[10]Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems.NewYork: ACM, 2017: 6000-6010.

【作者简介】

彭扬（1986—），女，重庆人，博士，讲师、硕士生导师，研究方向：机器学习算法及应用、脉冲控制及其应用、供应链管理、综合评价方法。

李昌翰（1994—），通信作者（email: 820974417@qq.com），男，四川南充人，硕士研究生，研究方向：机器学习、深度学习。

（收稿日期：2025-11-04 修回日期：2026-06-01）