EDMFormer：用于音乐结构分割的流派专属自监督学习

Sahal Sajeer滑铁卢大学
ssajeer@uwaterloo.ca

Krish Patel滑铁卢大学
k77patel@uwaterloo.ca

Oscar Chung滑铁卢大学
ohchung@uwaterloo.ca

Joel Song Bae滑铁卢大学
js2bae@uwaterloo.ca

arXiv:2603.08759v1 [cs.SD] 8 Mar 2026

摘要——**音乐结构分割是音频分析中的一项关键任务，但现有模型在电子舞曲（EDM）上的表现不佳。问题在于，大多数方法依赖歌词或和声相似性；这对流行音乐效果良好，却不适用于EDM。电子舞曲结构的定义依据则是能量、节奏和音色的变化，其中包含铺垫、Drop和分解段等不同乐段。我们推出EDMFormer，这是一种Transformer模型，利用EDM专属数据集和分类体系结合自监督音频嵌入。我们将该数据集发布为EDM-98，其中收录了98首经专业标注的EDM曲目。与现有模型相比，EDMFormer提升了边界检测和乐段标注的表现，尤其是在识别Drop和铺垫方面。结果表明，将学习到的表示与流派专属数据及结构先验相结合，对EDM十分有效，并且这一方法也可应用于其他特定音乐流派或更广泛的音频领域。**

I. 引言

音乐结构分析（MSA）是音乐信息检索（MIR）中的核心任务，可支持自动歌曲摘要、歌单生成、DJ辅助和音乐推荐等应用 [1]。现代音乐结构分析系统通常依赖自监督学习（SSL），从大规模音乐集合中创建音频嵌入 [2]。SongFormer 等基于Transformer的模型能够学习随时间变化的和声与旋律相似性模式，因此表现出色 [3]。

然而，这些模型假定音乐结构主要由和声重复和歌词分句定义，例如西方流行音乐中的主歌、副歌和桥段。这一假定并不适用于电子舞曲，因为其结构由不同特征驱动。电子舞曲中的铺垫和分解段并非由和弦或人声定义，而是由频谱亮度、节奏密度和低频能量变化来体现 [4]。因此，即使是顶尖的音乐结构分析模型也常会误判电子舞曲的段落，暴露出系统性的流派不匹配。

A. 研究动机

以往的 MIR 研究表明，当学习到的表示与目标流派的结构模式不匹配时，流派不匹配会降低模型性能 [5]。对于电子舞曲，切换点检测（SPD）研究表明，能量新颖度、鼓点起始次数和音色特征是识别结构转变的最佳线索 [6]。当前在流行音乐上训练的 SSL嵌入大多不包含这些线索。研究还指出，将节奏和能量结构特征与学习到的嵌入结合，可使边界检测效果提升 3% 以上[6]。

这些观察表明，由于分类体系不完善且缺乏相关数据，基于自监督学习的音乐结构分析模型会遗漏能量驱动型音乐所需的重要信息。这一发现促使我们采用针对电子舞曲的方案，整合结构更合理的分类体系与数据。

B. 相关工作

早期的音乐结构分析方法依赖色度特征和新颖度曲线等手工特征来检测结构边界 [1]。随着深度学习兴起，卷积神经网络和循环神经网络以学习到的表示取代了特征工程。近年来，基于Transformer的模型使用自监督学习目标进行训练，例如 MuQ 和 MusicFM，使 SongFormer 等模型能够在流行音乐数据集上取得优异表现 [3], [7], [8]。

然而，表示学习能力的提升并不一定能带来最优的结构分割。当领域专用节奏线索含糊不清时，若干研究报告指出，仅自监督学习嵌入的表现较差。节拍特征研究表明，明确设置节奏分类体系能够显著提升边界检测性能，尤其是在与神经嵌入融合时 [6]。同样，Joung等人证明，领域专门化能够提升模型在 MIR 任务中的鲁棒性 [5]。

对于电子舞曲，SPD 研究指出，能量新颖度、鼓点起始次数和谱通量是结构变化最有力的指标 [6]。但是，现有音乐结构分析研究尚未将这些发现纳入领域特定结构分割分类体系。此外，现有数据集（如 SALAMI）仍然几乎不包含电子舞曲，并采用以流行音乐为中心的分类体系 [9]。

这些空白凸显了以下需求：（1）电子舞曲数据集；（2）EDM专用结构分类体系；以及（3）将两者结合起来的自监督学习Transformer模型。

C. 问题定义

给定音频波形 X，音乐结构分割的目标是预测表示结构边界的一系列时间戳 $T = \{t_1, t_2, ..., t_n\}$，以及相应的段落标签 $Y = \{y_1, y_2, ..., y_n\}$。

现有方法通常会学习一种通用映射

$$f_{general} : E_{ssl}(X) \xrightarrow{\text{pop-centric dataset + general taxonomy}} (T, Y), \tag{1}$$

其中 $E_{ssl}(X)$ 是从预训练的自监督音乐模型中提取的嵌入，这些模型是在多样化的音乐语料库上训练的。

相比之下，我们的方法定义了一种领域专用映射

$$f_{edm} : E_{ssl}(X) \xrightarrow{\text{EDM-98 dataset + EDM taxonomy}} (T, Y), \quad (2)$$

该方法使用相同的 SSL 嵌入，但训练时利用 EDM-98 数据集及其 EDM 专用结构分类体系，使模型能够学习准确的模式和段落标签。

II. 方法

A. EDM-98 数据集

我们以其源语料库 EDM-CUE（包含 4710 首歌曲）为基础，整理构建了 EDM-98 数据集，其中包含 98 首 EDM 曲目，涵盖五种速度类别（120-180 BPM） [10]。两位专业标注人员以 ±0.5 秒的精度标注了结构段落（前奏、铺垫、Drop、分解段、尾奏、静音段和结束段）。该数据集采用分层、等额分配的子集构建方式，以应对 EDM-CUE 中严重的速度不平衡问题。

在具体实现中，我们提取每首曲目的 BPM（排除缺少 BPM 信息的曲目），随后将其分配到固定分桶中。样本尽可能均匀地分配到各个分桶，并通过剩余容量处理达到 N=98。每层内的样本均采用无放回随机抽样。

最终得到的 EDM-98 边际速度分布有意在中/高速度层之间保持接近均匀（House/Deep House 占 19.4%；Electro/Dance、Techno、Trance/Progressive、Dubstep/Bass、Drum & Bass/Jungle 各占 14.3%；Chill/Downtempo 占 9.2%）。这一分布带来了较大的 BPM跨度（均值 137.31，中位数 135.00，标准差 20.04，范围为 100–175）。

相比之下，完整的 EDM-CUE 数据集明显偏向高速度的 Drum & Bass/Jungle；其中训练划分（N=4710）的 BPM 均值为 169.52，中位数为 174.0，标准差为 11.62。所有曲目中有 90.5% 属于这一子流派 [10]。这种严重的类别不平衡导致其边际速度分布与 EDM-98 显著不同；EDM-98 有意拉平速度直方图，以支持模型在更广泛的速度/流派范围内进行稳健训练和评估。因此，EDM-98 应被视为一个经过精心构建的均衡子集，而不是 EDM-CUE 按比例缩小后的缩影。

B. 电子舞曲专用分类体系

EDM-98 中的所有曲目均使用旨在反映能量变化的电子舞曲专用分类体系进行标注。由于传统的流行导向分类体系（主歌、副歌、桥段）无法捕捉电子舞曲的结构，我们定义了一套新的 EDM 分类体系，包含以下类别：

- **前奏：** 曲目开头，能量较低，编曲较为稀疏。
- **铺垫：** 能量和张力逐渐增强，通常伴随着逐渐增强的鼓点模式或上升音效。
- **Drop：** 能量峰值段，包含主要节奏动机和低音线。
- **分解段：** 能量降低的段落，通常具有旋律性或氛围感，为 Drop 提供对比。
- **尾奏：** 曲目结尾部分，能量逐渐减弱。
- **静音段：** 几乎没有或完全没有音频信号的段落。
- **结束段：** 明确的曲目结束或最后一个音符。

图 1 展示了一首按照我们提出的分类体系进行标注的曲目示例，突出了电子舞曲的结构转变。

C. 模型架构

我们的模型架构大体沿用了 SongFormer 的设计，主要区别在于引入了电子舞曲专用分类体系、使用 EDM-98 进行训练的阶段，以及最后在 EDM-98 [3] 上进行的微调阶段。

我们从两个预训练音频基础模型 MuQ 和 MusicFM[7], [8] 中提取了自监督音乐表征。选择这两个模型作为基础模型，是因为它们都能从规模相对较大的音频语料库中生成时间信息丰富的嵌入，提供强大的低层声学表征。然而，至关重要的是，这两个模型的预训练都没有针对电子舞曲结构进行优化。它们学习到的表示反映的是一般声学模式，而非定义电子舞曲各段落的能量和音色变化。正因这种通用性，我们才有必要在数据集 EDM-98 上进行领域专用微调。EDMFormer 借助电子舞曲专用框架重新利用这些通用嵌入，教会模型识别对该音乐流派具有结构意义的声学模式。

我们融合了这两个模型，因为它们的预训练目标有所不同：MuQ 使用梅尔量化频谱目标进行掩码预测，侧重于音色和频谱特征；而 MusicFM 使用随机投影进行掩码令牌建模，并且由于其预训练期间采用了 30 秒的输入上下文，在较长范围的结构任务上展现出更强的性能 [7], [8]。我们结合这两种方法，为下游 Transformer 提供了互补信息。每个模型都生成了 1024 维嵌入，分别在两种时间上下文（30 秒和 420 秒窗口）下计算，因此每首曲目有四条嵌入流。随后，将这四个嵌入拼接，形成 4096 维表示。接下来，通过线性层将拼接后的 4096 维表示投影至 2048 维，并作为 SongFormer 编码器的输入。

图 2 更详细地展示了 SongFormer 的模型架构，以供参考。

D. 基线

我们将 EDMFormer 与 SongFormer（2025 年）的最新版本进行比较，将其作为主要基线 [3]。SongFormer 是当前表现最佳的基于自监督学习的音乐结构分析模型，在 SALAMI 等流行音乐基准测试中展现出优异的性能 [9]。该模型利用异构监督，在多样化的音乐语料库上进行预训练，并采用通用的结构分类体系（例如前奏、主歌、副歌、桥段、尾奏）。

SongFormer 可作为理想基线，原因有二。首先，它与 EDMFormer 共享相同的基础，从而能够

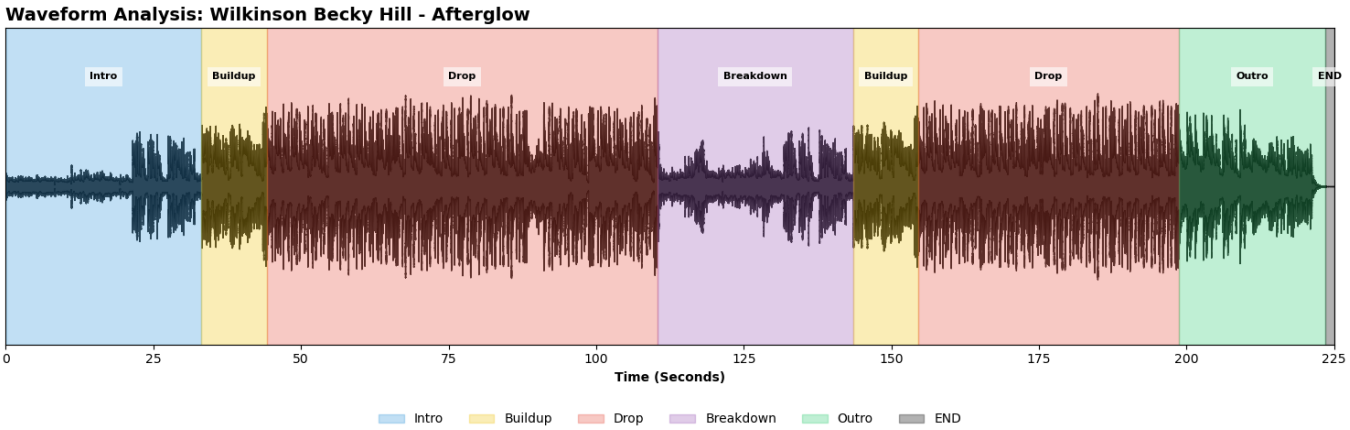

图1. 使用提出的电子舞曲分类体系标注的曲目片段示例。

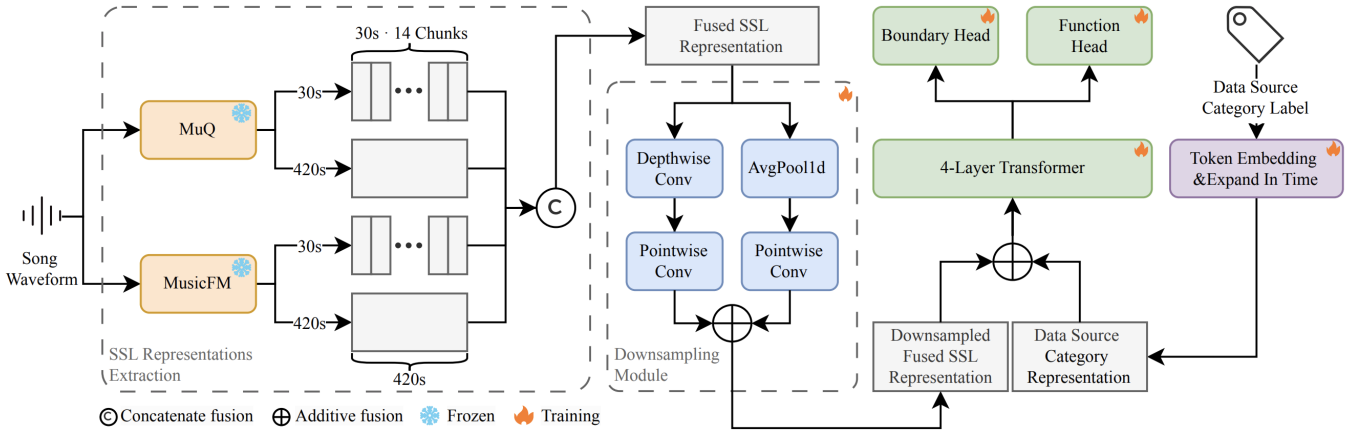

图2. SongFormer 的模型架构。

我们分离出领域特定数据与分类体系所带来的改进。其次，SongFormer 体现了当前衡量任何音乐结构分析模型所应参照的标准。我们在零样本迁移设定下评估了 SongFormer，将其直接应用于来自 EDM-98 的曲目且未进行微调，以检验通用音乐结构分析模型分析电子舞曲的效果。

E. 评估

我们使用命中率（HR）评估结果，该指标衡量预测边界落在真实边界容差窗口内的百分比。具体指标包括 HR@0.5s、HR@3s，以及逐帧标签准确率（ACC）；在电子舞曲数据集上采用 88-10 的训练测试划分。我们在训练集上进行了 5 折交叉验证，以进行超参数调优。

III. 结果

表I比较了我们提出的模型EDMFormer与业界基线SongFormer的性能（四舍五入至小数点后三位）。

表I EDMFormer 与 SongFormer 的性能

模型	HR@0.5	HR@3	ACC
SongFormer（流行音乐分类体系）	0.569	0.608	0.148
EDMFormer（EDM分类体系）	0.616	0.635	0.883

如表I所示，EDMFormer在三项评估指标上均优于SongFormer，在HR@0.5、HR@3和ACC上的提升幅度分别为4.7%、2.7%和73.5%。逐帧准确率（ACC）的提升最为显著：EDMFormer达到0.883，而SongFormer仅为0.148，提升了73.5%；这反映出以流行音乐为中心的分类体系与电子舞曲结构模式之间存在根本性不匹配。边界检测性能的提升进一步证实，EDMFormer能够学习到时间精度更高的转场线索，例如Drop起始点和Buildup峰值，而以流行音乐为中心的SongFormer则系统性地遗漏了这些特征。

三项指标上均取得的显著提升，支持了这一核心假设：类型不匹配是导致电子舞曲音乐结构分析表现不佳的主要因素。EDMFormer 的

架构在将融合后的表示传递给 SongFormer编码器之前，在两个时间尺度上融合 MuQ 和 MusicFM 的嵌入，既保留了通用 SSL嵌入的表达能力，又使 Transformer 能够针对由能量驱动的结构转变专门调整其注意力模式。所有指标和数据折叠上提升方向一致，这表明所提出的领域特定数据、分类体系与基于 SSL 的架构相结合，是一种有效且可复现的方法，可将音乐结构分析适配于特定音乐类型。

IV. 结论

本研究推出了 EDM-98 数据集，提出了 EDM 专用结构分类体系，并证明采用 Transformer 模型的自监督音乐表征能够提升电子舞曲的边界检测和乐段标注准确率。这些结果表明，将特定流派数据集和分类体系与基于 SSL 的 Transformer 模型相结合，是使音乐结构分析适用于专业流派的有效方法。

当前研究受到若干局限的制约。EDM-98 仅包含 98 首曲目。虽然这一数量足以验证概念，但它仅代表更大规模 EDM-CUE 语料库中的一小部分，而数据集的扩充又受限于高精度人工标注所需的时间。另一项局限在于基础模型本身：尽管 MuQ 和 MusicFM 能够提供强大的通用声学表征，但其预训练目标并未针对电子舞曲以能量为驱动的结构逻辑进行优化，可能无法充分捕捉音色和动态转变。最后，EDM-98 中的每首曲目均由一位标注者标注，因此可能引入偏差。

未来的工作可通过扩展数据集、整合显式声学线索以及采用标注者间一致性指标来弥补这些不足。未来工作还可纳入混合特征学习、层次化标注、适用于长曲目的相对位置嵌入，以及面向DJ的应用的实时推理能力。落实这些思路将提升未来迭代版本的可靠性和表征能力，并为长期开展领域特定的 SSL 预训练，以及面向DJ的应用实现实时推理奠定基础。

五、致谢

感谢各位标注人员为开发 EDM-98 数据集所作的贡献，也感谢 SongFormer、MuQ、MusicFM 和 librosa 背后的开源社区。本研究得益于滑铁卢大学人工智能团队 WAT.ai 提供的指导和计算资源。我们也感谢加拿大本科生人工智能会议（CUCAI）提供展示这项工作的机会。

参考文献

[1] J. Paulus、A. Klapuri等，《报告：基于音频的音乐结构分析领域的最新进展》，*ISMIR*，2010。[2] S. Liu 等，《音频自监督学习：综述》，2022。[3] C. Hao 等，《SongFormer：通过异构监督扩展音乐结构分析》，2025。[4] H. T. Zeiner-Henriksen，《“poumtchak”模式：电子舞曲中节奏、声音与动作之间的对应关系》，奥斯陆大学，2010。[5] H. Joung 和 K. Lee，《通过领域对抗训练学习稳健的音乐表示，实现音乐自动标签生成》，*ICASSP*，2024。[6] M. Zehren 等，《电子舞曲切换点检测方法的可解释性》，2024。[7] H. Zhu 等，《MuQ：基于梅尔残差向量量化的音乐自监督表示学习》，*IEEE*音频、语音与语言处理汇刊，2025。[8] M. Won、Y.-N. Hung 和 D. Le，《音乐信息学基础模型》，*arXiv*预印本 *arXiv:2311.03318*，2023。[9] D. De Roure、J. S. Downie 和 I. Fujinaga，《大规模音乐信息的结构分析（SALAMI）》，国际音乐信息检索学会会议论文集，2011。[10] G. Arg¨uello、L. Lanzend¨orfer 和 R. Wattenhofer，《使用目标检测估计提示点》，*arXiv*预印本 *arXiv:2407.06823v1*，2024。