arXiv:2303.01884v1 [cs.SD] 3 Mar 2023

AutoMatch：用于提升深度学习视频剪辑助手的音频节拍匹配基准

SenPei[1,2]，JingyaYu[1]，QiChen[1]，WozhouHe[11]字节跳动[2]
NLPR，中国科学院自动化研究所{peisen, yujingya.yjy,
chenqi.02, hewozhou}@bytedance.com

摘要

短视频的爆发式增长彻底重塑了人们的社交方式，催生了日常分享和获取最新信息的新趋势。这些丰富的视频资源一方面得益于配备摄像头的便携设备的普及，另一方面也离不开众多视频创作者贡献的宝贵剪辑工作。在本文中，我们研究了一个新颖且实用的问题，即音频节拍匹配（*ABM*），旨在根据背景音乐推荐合适的转场时间戳。该技术有助于减轻视频剪辑过程中的繁重劳动，为创作者节省精力，使其能够更加专注于视频内容的创意。我们正式定义了*ABM*问题及其评估协议。同时，我们提出了大规模音频数据集*AutoMatch*，其中包含超过*8.7*万首经过精细标注的背景音乐，以推动这一新兴研究方向的发展。为了进一步为后续研究奠定坚实基础，我们还提出了一种名为*BeatX*的新模型，用于解决这一具有挑战性的任务。此外，我们创新性地提出了标签范围这一概念，一举消除了数据不平衡问题，并在训练过程中为真实值分配自适应权重。尽管短视频平台已经蓬勃发展了很长时间，但针对这一场景的相关研究仍然不足；据我们所知，*AutoMatch*是首个用于解决音频节拍匹配问题的大规模数据集。我们希望，所发布的数据集和具有竞争力的基线能够引起人们对这一研究方向更多的关注。数据集和代码将公开发布。

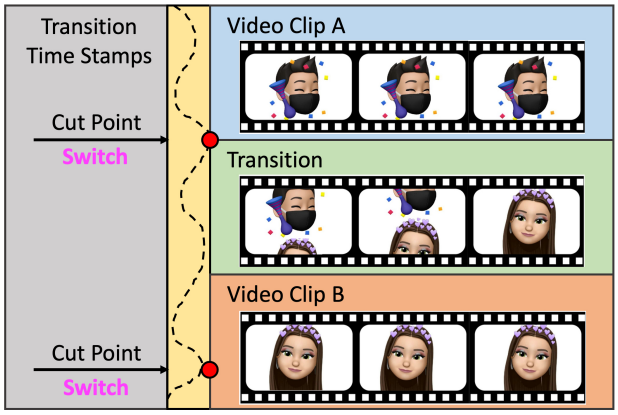


图 1. 手动视频剪辑示例。上方每组图像展示了视频中出现的一个场景。左侧的波形表示相应的音频信息。视频创作者在不同的视频情境之间添加转场效果，并将转场放置在恰当的位置，以匹配音频节拍。如果不使用预定义模板，用户就必须费力地定位音频的切分点，这对视频创意的提升作用有限。

1. 引言

深度学习研究为人们的生活带来了切实的便利，例如生物信息认证 [30, 34, 32, 57, 12, 23, 38], 推荐系统 [6, 52, 16, 43, 49, 31, 5, 7, 29], 和辅助机器人 [21, 54, 26, 56, 11, 3, 1], 等。不论任务是宏观还是具体，解决一个问题往往会带动相关应用的爆发式增长，使人们免于繁重、重复且不需要专业技能的操作。受到许多与大数据和人工智能相关的研究启发，例如 [48, 50, 35, 36, 33], ，我们认为，在大数据时代，恰当地定义或建模一项实际任务，在某种程度上应与解决问题同等重要，因为这是推动向善人工智能的有效途径。这一原则促使众多人工智能研究者探索并涉足那些通常不受关注的领域，让人工智能研究的成果惠及更广泛的人群。

在本文中，我们秉持技术应服务于生活的初衷，因此聚焦于一个意义重大但相关研究较为匮乏的领域，即检测音频中的潜在切分点。

通常，剪辑点表示视频中建议更改内容的位置，且通常会插入转场，以连接不同的上下文。我们将这一任务称为音频节拍匹配（ABM）。该任务在视频创作中发挥着至关重要的作用，因为视频内容切换与背景音乐节奏之间的不匹配，会给观众带来不适感。目前，在视频剪辑过程中，编辑人员大多仍需手动完成上述音频节拍匹配任务，这耗费了大量时间和精力。为了赋能视频创作者并推动视频剪辑自动化，我们注意到并进一步研究了这一鲜有人关注的方向，即 ABM，旨在使深度神经网络具备类似人类的剪辑能力。

据我们所知，现有研究已利用过零率等方法，开发出检测音频变化边界的能力 [8]以及经典的傅里叶变换。然而，这些能力对视频创作者的帮助仍然有限。例如，在许多视频中，我们很容易发现，音乐的变化边界未必会引发转场。这表明，前述切分点的时间戳虽与音乐变化密切相关，却并不完全取决于音乐的局部变化；这也说明，视频剪辑师必须考虑音乐的局部边界之外的因素，例如整首音乐的风格。

为了帮助深度模型培养类似人类的剪辑能力，我们收集了首个大规模**音频到转场**数据集，即 *AutoMatch*，其中包含超过 8.7 万首背景音乐及其对应的切分点时间戳。在此基础上，我们进一步开发了一种名为*BeatX* 的算法，可利用局部到全局特征自动分析原始音频文件，并给出推荐的转场位置。这项辅助技术将为视频剪辑提供一些前期支持，丰富视频创作的风格。本文的主要贡献可概括为以下三个方面：

- 我们正式定义了音频节拍匹配（ABM）任务及其评估协议，旨在帮助视频创作者根据音频信息定位合适的转场时间戳。这项技术将为众多视频剪辑师提供推荐的切分点，帮助他们部分摆脱繁琐的预处理工作。

- 我们收集了首个名为*Au-toMatch* 的大规模数据集，以推动 ABM 研究。*Au-toMatch* 包含超过 8.7 万首经过精细标注的背景音乐及其对应的转场时间戳。该数据集使 ABM 得以受益于深度学习方案。

- 我们提出了*BeatX*，该方法充分利用经典技术和深度学习技术，实现了具有竞争力的检测性能。此外，我们还提出了标签范围，从而消除了数据不平衡问题，并为损失分配自适应权重。大量实验和详细的消融实验展示了我们所提出的*BeatX*的见解与有效性。我们希望本研究能够激励更多研究人员关注那些不热门但实用的方向，利用最新技术提升生活的便利性。

2. 相关工作

本节简要概述与本文最相关的研究，包括音频表示学习、自动视频剪辑以及通用事件边界检测（GEBD）。

音频表示学习。 无论是在手工设计描述符时代，还是在深度特征时代，表征的设计都决定着模型的性能。在 [9],这项工作中，研究者综合了多种经典音频特征，例如梅尔倒谱 [53], 音高类别轮廓 [10], 过零率 [28, 55], 和短时能量 [9]。所有这些特征都是经过精心人工设计的，并具有较强的泛化能力。 [41] 进一步指出了音频特征演变的四个阶段，分别是时域、频域、时频域集成，以及最新的基于深度学习的特征。本文同时采用经典的人工设计特征和学习得到的深度表示，并报告它们各自的效果。

自动视频剪辑。 这项技术为视频剪辑师完成了大量预处理工作，是基于深度学习的辅助工具的代表之一。Auto-Transition [44]旨在根据两个连续视频片段的视觉和音频信息，推荐它们之间的转场。这项工作还正式定义了视频转场推荐（VTR）任务及其相应的评估协议，降低了非专业人士进行转场选择的难度。非线性剪辑迁移 [13] 研究如何将源视频的剪辑方式应用于目标视频。Text2live [2] 开发了使用输入文本编辑图像或视频的能力，降低了视频剪辑的门槛。MovieCuts [37]对多模态信息进行建模，以识别剪切类型，为视频剪辑行业带来了全新的体验。CMVE [25] 利用视觉和文本元数据，还能突出原始视频中与主线故事相关的关键对象，帮助剪辑师快速定位他们关注的片段。谈话头像视频剪辑 [14]致力于帮助剪辑师更改采访视频中的语音内容或删除填充词。以上我们仅列举了深度学习视频剪辑辅助工具的几个代表性实例。这些方法要么解决了一些实际问题，减轻了视频创作者的工作负担，要么开辟了尚未得到充分探索的新研究方向。

受这些启发，在本工作中，我们关注一个不太热门但实用的方向——音频节拍匹配（ABM）。它旨在依据背景音乐寻找潜在的转场位置，从而减轻剪辑师此前需要承担的部分工作。

通用事件边界检测。GEBD 长期以来致力于在原始视频中定位不依赖分类体系的事件边界，并将视频切分为若干连续的时间片段。相关技术 [18, 20, 39, 46] 将固定长度的一组帧构建为上下文，并分别判断每一帧是否属于边界。与这些现有方法不同，文献 [27, 19]中提出的 SC-Transformer 引入了 SPoS 模块，可高效地为每一帧生成上下文，并借助 Transformer 编码器进行时间融合，从而增强沿时间轴的特征交互。对于音频节拍匹配（ABM）任务，转场的位置也可视为背景音乐的边界。然而，与可利用局部上下文识别的 GEBD 不同，ABM 中出现的边界并非仅由局部节奏（即上下文）决定，这意味着背景音乐的全局信息会显著影响推荐性能。

3. 预备知识

本节将详细介绍所提出的音频节拍匹配（ABM）任务的背景，包括问题定义和评估协议。

3.1. 问题定义

给定如 x 所示的完整音频输入，我们以基本时间单位将其划分为若干个连续的片段，即 $x = \{x_1, x_2, ..., x_N\}$。该时间单位表示每个音频片段 x_k 的持续时长。鉴于原始音频 x 的时长可能无法被该时间单位整除，我们在 x 的末尾使用零进行填充，这意味着所有音频片段的总时长不短于原始输入 x 的时长。相应地， x 的目标标签表示为 y，它是一个包含 N 个值的二元向量。例如，$y_i = 1$ 表明，建议在第 i 个时间单位之后更改视频内容并分配适当的转场。ABM 任务旨在构建一个预测器 $f(\cdot)$，该预测器以音频 x 作为输入，并输出一个 N 维的预测向量 $\hat{y}$。预期预测结果 $\hat{y}$ 应尽可能类似于真实值 y。

3.2. 评估协议

我们使用召回率（R）和精确率（P）来评估 ABM 中预测器的性能。首先，我们给出*hit@k*的定义。如果预测 $\hat{y}_i$ **命中** 目标视频 y，且 $\hat{y}_i$ 和 y_i 均等于 1，我们就称其为命中，这意味着应在第 i 个 时间单位之后插入转场。此外，我们放宽评估标准。如果预测 $\hat{y}_i$ 与$\{y_{i-k}, y_{i-k+1}, ..., y_i, ..., y_{i+k-1}, y_{i+k}\}$之间的某个目标视频均等于 1，我们就将这种情况称为*hit@k*。显然，传统的预测性能是在*hit@0*的设置下计算的。基于这些符号和定义，我们构建$P@k$（精确率）和$R@k$（召回率），如下所示：

$$P@k = \frac{hit@k(y, \hat{y}_i)}{\sum_{i=1}^{N} \hat{y}_i} \quad R@k = \frac{hit@k(y, \hat{y}_i)}{\sum_{i=1}^{N} y_i} \tag{1}$$

其中，*hit@k(y,ˆy)* 表示在 y 和 $\hat{y}$ 之间，按照 *hit@k* 的设定计算出的重叠值数量。基于公式1中的上述协议，可以轻松计算出 F1分数、AUPR 和 AUROC 等其他指标。

4. *AutoMatch*数据集

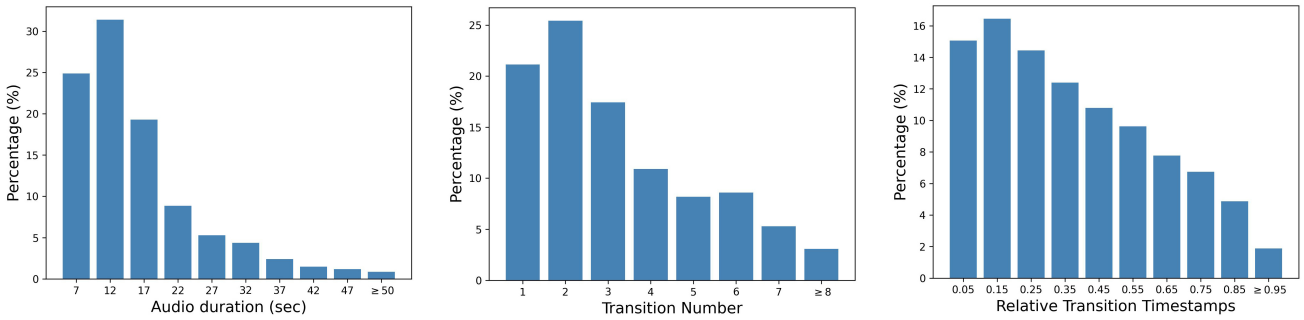
本节将详细介绍数据收集、质量控制和数据安全的流程。随后，我们将展示*AutoMatch*的统计信息。请注意，音频提取自收集到的在线视频。

4.1. 数据收集

收集如此大规模的数据集无疑是一项艰巨的任务。为了提高效率并避免繁琐的审批流程，我们规定，只有满足以下规则的视频才能作为*AutoMatch*的候选视频：1). 视频在平台上开源，所有用户均可公开访问，且视频创作者允许互联网用户基于其原始作品进行剪辑或再创作。这有助于我们方便地获得视频上传者/所有者的完整授权；2). 视频中使用的内容，包括但不限于图像、主题和音频等，必须健康积极，即不得包含色情画面、令人不适的音乐或个人信息等。这一限制对于确保*AutoMatch*不涉及伦理顾虑或问题至关重要；3). 不存在版权问题。视频中使用的背景音乐不得侵犯他人的版权。由于短视频分享平台的审核机制严格，这一点相对容易保证。对于符合上述标准的视频，我们会将其背景音乐选为候选音乐；此外，我们还会利用视频创作者标注的标签，检查视频中的转场是否与音乐高度相关，以保留节奏感强的视频并排除噪声数据，即转场与音频的相关性很低。

4.2. 数据统计

我们分析了音频时长、每段音频中的转场数量，以及各转场对应的时间戳分布。统计结果如图2所示。从图2 (a)中可以看出



(a) (b) (c)

图 2. 所提出的*AutoMatch*的统计结果。(a) 展示了音频时长的分布。(b) 展示了每段背景音乐中的转场数量（即切分点）。可以看出，大多数背景音乐对应的转场少于 8 个。(c) 展示了各转场在对应视频中出现的相对时间戳。显然，相对转场时间戳的分布较为均匀，并未集中在某些特定区域，这意味着*AutoMatch* 的标注丰富多样，因此难以学习。这些复杂的真实标注有助于验证模型是否学到了有效的表征，还是仅仅记住了该特定数据集中的模式。

如果一段音频超过一分钟，但对应的剪辑点少于十个，则会被排除在超长音频之外。这些音频通常旋律非常舒缓，对应的剪辑点在时间上分布均匀，因此对学习判别性特征的贡献较小。图2 (b)展示了每段音频中剪辑点数量的分布。由于我们在*AutoMatch*中排除了超长且节奏平缓的音乐，大多数音频中的转场/剪辑点都少于八个。图2(c)显示，剪辑点在音频中均匀分布，从而避免了深度神经网络直接捕捉该数据集特定模式的捷径学习 [15, 40]。

4.3. 数据安全

我们通过两个步骤来确保数据安全。首先，我们在数据收集过程中引入了检查机制（参见第 4.1 节）。此外，我们还与内部法务/隐私工作人员合作，确保研究过程中的操作符合相关法律要求，包括数据收集、数据处理和数据使用等。我们的研究不雇用工作人员，也不招募人类参与者。因此，我们无需提前准备知情同意书（ICF）。*AutoMatch* 中使用的所有数据均已由内部法务部门审核，从而避免了潜在的法律/伦理风险和版权问题。

5. 方法

本节详细介绍模型架构设计的动机、损失函数的构建以及复现流程。

5.1. 架构设计

在模型设计中，我们聚焦于三个基本问题：1）。使用人工设计特征或自适应学习的表征。

如第2节所述，深度学习模型可以借助精心设计的描述符（如梅尔频谱 [53] 和短时能量 [9], 得到增强；在设计中我们采纳了这一建议，对传统特征与深度特征进行融合；2）。如何捕捉相邻音频之间的关系，即相邻的音频片段。该问题的主要挑战体现在两个方面：选择合适的窗口大小以执行局部注意力，并在训练过程中高效地为每个片段生成结构上下文 [27] ；3）。如何以相对于音频时长的线性复杂度执行全局注意力。正如许多研究中所观察到的，
[4, 42, 45, 51, 47, 24, 22],尽管 Transformer 在许多复杂任务上取得了令人印象深刻的性能，但由于其复杂度为 $O(N^2)$（相对于序列长度）N，将其训练或部署到实际场景中的成本高得令人难以承受。由于 ABM 服务于实际的实时场景，推理成本显得尤为关键。

总而言之，我们认为，适用于音频节拍匹配任务的模型应采用具有良好泛化能力的音频表示，高效获取结构上下文，并以相对于音频时长近似线性的复杂度，在整个音频上执行全局注意力。为应对上述挑战，我们设计了*BeatX*，该模型同时将原始音频信号和手工设计的描述符作为输入，并借助 SPoS [27]快速生成结构上下文，同时使用训练和部署成本均较低的 LinFormer [51]融合局部特征。我们提出的*BeatX* 整体框架详见图3 。为便于描述，如上所述，我们将 *BeatX* 划分为三个组成部分：时间维度特征提取（TFE）、局部上下文生成（LCG）和线性全局融合（LGF）。按照图3所示的流程，我们对各个模块进行形式化描述，并定义构建损失函数所需的符号

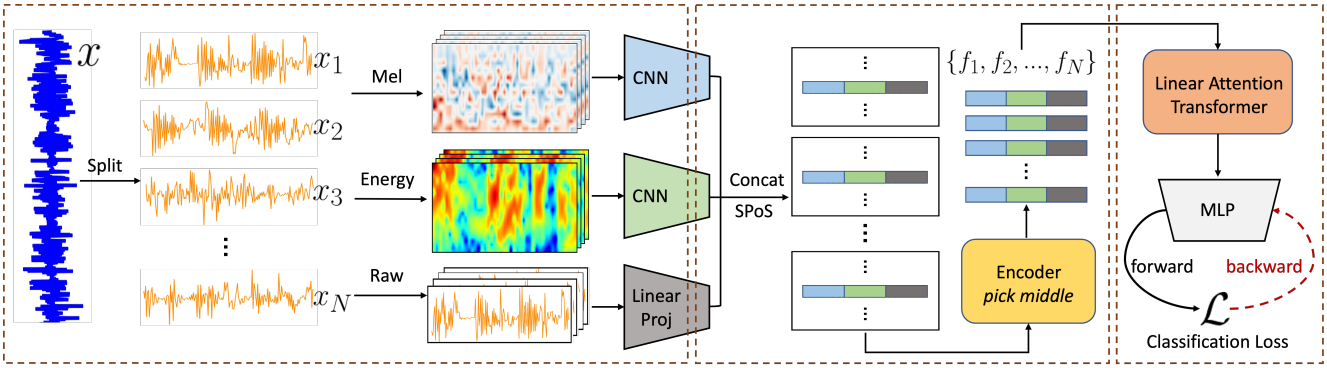


图 3. 所提出的 *BeatX* 框架。在上图中，我们使用虚线框突出显示了前述 *BeatX* 的三个组件。从左到右依次为时间维度特征提取（TFE）、局部上下文生成（LCG）和线性全局融合（LGF）。输入信号被划分为连续的音频片段，每个片段的时长等于预定义的时间单位。梅尔倒谱和短时能量被用作人工描述符，同时，原始音频信号将经过线性投影层处理。所有这些特征会拼接为中间表示。我们采用 SPoS [27] 来生成每个音频单元的结构上下文 x_k。随后，常规编码器和线性注意力变换器（LinFormer）[51]负责完成局部和全局特征融合。整个流程被构建为一个分类问题，并使用反向传播技术进行优化。

用于构建损失函数的选项。

时间维度特征提取（TFE）。如第 3.1 节所示，给定原始音频输入 x，我们首先将其划分为N个连续的音频片段，即 $\{x_1, x_2, x_3, ..., x_N\}$。每个音频片段的时长等于预定义的时间单位。这些子片段之间没有重叠。为了利用人工设计描述符中的语义信息，我们将梅尔倒谱和短时能量与原始信号一同输入。上图中的 CNN 是 ResNet-18 [17]的前两个模块。原始信号经过由 ReLU 激活的线性层投影。TFE 模块的输出是上述特征的拼接，即 $g_i = M(x_i) \circ E(x_i) \circ L(x_i)$，其中 $M(\cdot)$、$E(\cdot)$ 和 $L(\cdot)$ 分别表示梅尔倒谱、短时能量和线性投影。我们使用 $\circ$ 表示拼接操作。TFE 中不考虑局部特征之间的时间交互。

局部上下文生成（LCG）。该模块旨在为每个时间单位（即，x_i）高效生成结构上下文。给定输入 $x = \{x_1, x_2, x_3, ..., x_N\}$和局部窗口大小 k，LCG 会生成长度为 $2k+1$的 x_i 上下文，表示为 $\{x_{i-k}, x_{i-k+1}, ..., x_i, ..., x_{i+k-1}, x_{i+k}\}$。对于时长不足的上下文，例如最前和最后的 k 个片段，会用零向量填充。假设分割后的音频表示为 $\{x_1, x_2, x_3, ..., x_N\} \in \mathbb{R}^{N \times C}$，其中 C 为通道维度，那么 LCG 的结构化输出为 $\{c_1, c_2, c_3, ..., c_N\} \in \mathbb{R}^{N \times k \times C}$，其中 $c_i \in \mathbb{R}^{k \times C}$ 表示音频的局部上下文 x_i。SPoS [27] 通过精心设计的数组排列满足上述要求。由于这更像是一种工程方案，此处不作详细介绍。为联合关注不同时间戳的信息，每个结构上下文 c_i 都由多头编码器融合。我们只选取每个上下文 c_i的中间词元，输出记为 $f_i \in \mathbb{R}^C$。结构化特征组 $\{f_1, f_2, f_3, ..., f_N\}$ 分布在$\mathbb{R}^{N \times C}$的空间中，这与 $\{x_1, x_2, x_3, ..., x_N\}$相同。

线性全局融合 (LGF)。通常，Transformer架构的主要效率瓶颈来自自注意力机制，因为每个词元的表示都必须与前一层中的所有其他词元计算相关性。简而言之，给定 Q、K 和 V（$\mathbb{R}^{N \times d}$）作为输入，自注意力旨在学习一组权重 W^Q、W^K 和 W^V（$\mathbb{R}^{d \times d}$），以保留序列中的长期信息，输出构造如下：

$$h = \texttt{softmax}[\frac{QW^Q(KW^K)^T}{\sqrt{d}}]VW^V = PVW^V \quad (2)$$

其中，d 是由 W^K 生成的投影空间的隐藏维度，而 P 是等效的上下文映射矩阵。我们不区分 W^Q、W^K 和 W^V 映射前后的特征维度。计算 P 在时间和空间上的复杂度均为 $O(N^2)$，因为 Q、K 和 V 的序列长度均为 N。LinFormer [51] 指出并验证，自注意力的映射矩阵 P是低秩的，因此可通过线性投影分解为多个较小的注意力。例如，如果我们选取 E 和 F（$\mathbb{R}^{p \times N}$、$p \ll N$）作为作用于 K 和 V的线性投影，则可如下构造近似的 h：

$$h \approx \hat{h} = \texttt{softmax}[\frac{QW^Q(E \cdot KW^K)^T}{\sqrt{d}}]F \cdot VW^V \quad (3)$$

注意，p 是线性投影的中间维度。在上述公式中，QW^Q 与 $(E \cdot KW^Q)^T$相乘的复杂度仅为 $O(N)$，因为p 和 d 都是常数。这一操作显著降低了自注意力的复杂度，从 $O(N^2)$降至 $O(N)$，

这一原理 $p \ll N$被我们用于设计全局融合模块。引入的性能下降通过 LinFormer [51]进行评估。

5.2. 损失函数

本节着重解决两个关键问题：1). 如何平衡正负样本，因为只有极少数时间戳被标注为正确的转场位置，这意味着负样本数据远多于正样本；2). 如何衡量模型预测所产生的损失的重要性，因为预测存在较小的时间偏移（例如，仅一个时间单位）在一定程度上是可以接受的。为简便起见，输入 x 的标签表示为 $y = \{y_1, y_2, y_3, ..., y_N\}$，因为 x 被划分为 N 个连续片段/时间单位。 $y_i = 1$ 表示应在第 i 个片段之后立即插入转场。

借助这些辅助符号，我们使用**标签范围** $s_i(k)$来解决上述问题，该范围用于衡量**正**标签 y_i 对第 k 个时间单位产生的影响。直观而言， y_i 的影响会随距离增加而减弱。例如，第一个时间单位的标签对最后一个时间单位的预测影响甚微。根据经验，标签范围 $s_i(k)$构造为一个高斯函数:

$$s_i(k) = \exp[-\frac{(i-k)^2}{2\sigma^2}] \tag{4}$$

其中 σ 为高斯分布的半径。根据公式 5，我们解决了本节开头提出的第二个问题。此外，我们使用超参数 σ 来控制正样本数据与负样本数据之间的平衡。显然， σ 越大，越强调正样本数据，因为高斯分布沿时间轴方向拉伸得更宽。此外， σ取决于数据集，因为它应根据特定数据集的正负标签分布来确定。为了获得合适的 σ值，我们参考第 6.3.2 节中的消融实验。目前，我们选取$\hat{\sigma}$ 作为 σ的最合适拟合值，并将标签范围细化为 $s_i(k, \hat{\sigma})$。采用这些记号，给定输入 x，其中包含 N 个片段及其对应的目标视频 y，我们得到第 k 个时间单位的总体重要性如下:

$$S(k, \hat{\sigma}) = \frac{1}{Z} \sum_{i=1}^{N} s_i(k, \hat{\sigma}) \cdot y_i \tag{5}$$

其中 Z 是归一化因子，可确保重要性最大值不超过 1。 k 从 1 开始到 N。 $S(k, \hat{\sigma})$被视为标签范围掩码，用于平衡正样本/负样本数据，并衡量每个时间单位的重要性。假设 *BeatX*的预测结果如 $P = \{p_1, p_2, p_3, ..., p_N\}$所示，我们将完整损失目标的计算方式如下:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} S(i, \hat{\sigma}) \cdot [y_i \log(p_i) + (1-y_i) \log(1-p_i)] \tag{6}$$

具体而言，对于没有转场的视频，我们省略掩码 $S(k, \hat{\sigma})$，因为它始终等于零。

5.3. 复现声明

我们由衷赞赏计算机视觉社区所营造的开源氛围。本研究使用的详细代码和数据集将公开发布。此外，我们还提供脚本，帮助用户根据候选图像和原始音频渲染视频。我们设计了便于用户使用的界面，可将 MP3/MP4/MOV 文件作为输入，并自动提取音频信息。借助*BeatX*推荐的转场时间戳，渲染脚本将生成一个示例，整合音频、图像、转场和视频特效。我们的示例已附在**补充材料**中。

6. 实验

本节将详细介绍我们的实验，报告*BeatX*的比较结果，展示*BeatX*各组成部分的效果，并对超参数进行消融实验。

6.1. 实验设置

可比较方法。由于这一方向尚无充分的研究成果，与以往方法进行比较并非易事。为证明我们提出的 *BeatX*的有效性，我们设计了几种可比较的基线方法: Linear 线性分类器、一维卷积神经网络和基于自注意力的编码器。我们遵循的原则是，所有可比较方法的计算复杂度（即可训练参数数量）相近。全连接分类器（表1中的 Linear）由四个经 ReLU 激活的线性层构成。基于自注意力的编码器每层有 8 个注意力头，共 4 层。编码器的前馈网络（FFN）维度设为 512。卷积模型采用四个堆叠的一维卷积层，随后使用两个连续的全连接层对中间特征进行投影。在*BeatX*中，梅尔倒谱、短时能量和原始信号的特征维度均为 128。LCG 中使用的编码器包含 2 个注意力层，每层有 4 个注意力头。LGF 中仅使用一个包含 4 个注意力头的注意力层。

训练与推理流程详情。所有可比较的方法均训练 300 个周期，默认将时间单位设为 0.1 秒。学习率从 2e-4 开始，每 20 个周期减半。AdamW

方法	参数	*ACC*	*hit@0*			命中率 *@1*			命中率 *@2*		
			P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
Linear	5.44M	75.97	65.46	49.26	53.82	81.77	52.20	63.16	88.43	65.74	71.60
一维卷积神经网络	5.54M	75.86	67.41	46.38	53.50	83.88	48.38	60.13	87.04	64.52	70.93
编码器	5.64M	77.12	67.10	47.29	54.20	85.91	64.35	73.24	88.27	71.30	79.82
BeatX	5.10M	**79.95**$_{+2.83}$	67.21	60.36	**61.58**$_{+7.38}$	86.40	77.66	**79.11**$_{+5.87}$	91.23	80.76	**83.08**$_{+3.26}$

表1. 与先前研究的可比结果。 *hit@k* 的详细定义见第 3.2节。所示的整体指标通过对逐音频预测结果取平均获得。为保持训练复杂度相近，我们选取了可比的方法。值得注意的是，*hit@k* 也可以通过 SPoS 高效解决 [27], 我们在随附的补充材料代码中对此进行了详细说明。我们用粗体突出显示最佳性能，并用下划线标出紧随其后的方法。

被用作优化器。我们的实验未使用预训练权重，这意味着所有方案均从头开始训练。我们从数据集中划分出 7,000 个样本用于验证，并报告其最佳指标。总体性能通过对逐音频结果取平均获得。窗口大小（即半时长 k，即 LCG 中局部上下文的半时长）设为 5。所有实验均使用 8 张 NVIDIA V100 [1] GPU 进行。

6.2. 主要评估结果

表1 展示了音频节拍匹配的性能。如果某个时间单位对应的预测置信度高于 0.5，则将其判定为合适的过渡位置。我们在整个实验中均采用这一阈值。*hit@0* 如表 1 所示，是常规的评估协议。从实验结果可以清楚地看出，*BeatX* 显著优于 *Linear* 和 一维卷积神经网络，平均 F1分数提升超过 7.76%。这表明自注意力机制不可或缺，即，捕获相邻邻居特征的重要性。更具体地说，通过比较 *BeatX* 和 编码器，我们还可以得出结论：全局注意力能够进一步提升音频节拍匹配性能，因为 ABM 任务需要全局视角，检测增益约为 7.38%，同样相当显著。

6.3. 消融实验

本节针对以下问题进行分析：1). 如何确定最合适的高斯半径 $\hat{}\sigma$；2). 验证标签范围掩码的效果；3). 研究*BeatX*中不同特征组合的效果，即梅尔倒谱、短时能量和线性投影；4). 详细分析*BeatX*中超参数的消融实验，例如自注意力头的数量、自注意力层的深度，以及拼接特征的中间维度等。

[1]https://www.nvidia.com/en-sg/data-center/v100/

$\hat{\sigma}$	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3
正负样本比例	2.21	1.53	1.20	1.01	0.89	0.81	0.75	0.71

表2. 高斯半径$\hat{\sigma}$ 选择的消融实验。除非另有说明，我们将 0.9 设为$\hat{\sigma}$ 的默认值。

6.3.1 标准差$\hat{}\sigma$的估计

给定目标标签 y 和相应的标签范围掩码 S，我们使用正负样本比例（正样本/负样本）来描述正负样本是否均衡。形式上，我们按以下方式定义正负样本比例：

$$PN = \frac{\sum_{i=1}^{N} S(i,\hat{\sigma}) \cdot y_i}{\sum_{i=1}^{N} S(i,\hat{\sigma}) \cdot (1-y_i)} \tag{7}$$

我们的目标是选择合适的 $\hat{}\sigma$，使正负样本比例分布在 1 附近。为实现这一目标，我们计算正负样本比例在 $\hat{}\sigma$ 取值从 0.6 到 1.3、步长为 0.1 时的数值，消融实验结果详见表 2。需要注意的是，我们计算正负样本比例时，会统计 *AutoMatch* 中每个数据样本的比例，并报告其平均值。

6.3.2 标签范围掩码的影响 S

我们比较使用和不使用标签范围掩码时的推荐性能 S。详细结果见表 3。所有方法均在*hit@0*（即传统评估协议）下进行评估。我们分别用粗体和下划线标出使用和不使用标签范围时的 F1分数。下表所报告的性能提升即为两者之差。从表 3可以看出，标签范围掩码 S 至少带来了 3.06% 的提升，证明了长尾训练数据中数据均衡的必要性。每组实验运行两次，并取其平均性能。

6.3.3 不同特征组合的消融实验

我们进一步研究了手工设计的描述符和学习得到的自适应特征分别带来的影响。表4中，

方法	S	P	R	$F1$	ACC	$F1$增益
Linear	✗	64.00	37.48	46.01	74.15	/
	✓	65.46	49.26	**53.82**	75.97	↑ **7.81**
一维卷积神经网络	✗	66.99	39.43	47.22	75.04	/
	✓	67.41	46.38	**53.50**	75.86	↑ **6.28**
编码器	✗	66.90	40.45	48.97	74.55	/
	✓	67.10	47.29	**54.20**	77.12	↑ **5.23**
BeatX	✗	65.70	57.42	58.52	74.73	/
	✓	67.21	60.36	**61.58**	79.95	↑ **3.06**

表3. 对我们提出的标签范围掩码进行的消融实验 S。 ✗表示未执行标签范围处理，而 ✓ 表示该方法采用标签范围掩码 S 来生成自适应损失权重。

Mel	能量	*Raw*	P	R	$F1$	ACC
✓			69.83	53.34	57.37	77.85
	✓		66.46	58.96	58.12	77.92
		✓	69.42	47.18	55.06	76.33
✓	✓		68.80	53.06	59.16	78.65
	✓	✓	65.69	57.21	59.75	78.33
✓		✓	66.67	59.61	60.25	78.42
✓	✓	✓	67.21	60.36	**61.58**	79.95

表4。*BeatX*中特征组合的消融实验。我们用粗体和下划线标出性能最佳的前两个模型。双下划线表示人工描述符和深度特征中最具判别性的特征。F1分数反映了模型的整体性能。

其中，梅尔、能量和原始分别表示梅尔倒谱、短时能量和原始音频信号。所有实验均采用标签范围掩码进行数据平衡。从表4所示结果可以看出，仅使用原始信号输入时，检测性能（F1分数）为55.06%，低于使用梅尔倒谱和短时能量的组别。单独使用能量特征可获得58.12%的F1分数，比原始信号高出3.06个百分点。通过特征融合，*BeatX* 在验证数据上的表现更好，F1分数约为61.58%。我们认为，这是因为手工设计的描述符有助于防止模型进行捷径学习。

6.3.4 *BeatX*超参数的消融实验

BeatX 的关键参数包括投影维度、自注意力模块的深度以及注意力头的数量。考虑到*BeatX* 将部署于实际场景中，我们不会对过深的注意力架构进行消融实验。总体结果如下表所示。速度是在单个 NVIDIA

#Proj	#Head	#Depth	#Param	速度	$F1$
128	4	2	5.10M	1873.53	61.58
256	4	2	15.15M	1588.22	61.26
512	4	2	54.12M	1434.44	**62.03**
128	6	2	5.10M	1726.07	61.23
128	8	2	5.10M	1656.89	60.98
128	12	2	5.10M	1635.04	61.16
128	4	4	6.49M	1022.05	60.49
128	4	6	7.87M	724.40	59.87
128	4	8	9.25M	582.15	58.65

表5. 关于*BeatX*超参数的消融实验。速度表示模型每秒处理的时间单位（0.1秒）数量。选择模型架构时应同时考虑其性能和效率。

V100 GPU 上使用单精度浮点数数据类型进行评估的。由表 5中的结果可明显看出，随着注意力深度增加，推理速度会显著下降。例如，将注意力层数从 2 增加到 4（即翻倍）时，推理速度几乎减半，从每秒 1873.53 个时间单位降至 1022.05。相比之下，推理速度对注意力头数量和中间特征维度的变化不太敏感。随着投影维度增大，可训练参数的数量会急剧增加。默认情况下，我们使用包含 2 层和 4 个注意力头的局部注意力模块，即第一组。

6.3.5 局部窗口大小消融实验 k

我们针对半窗口大小进行了消融实验 k，该大小在第3.2节中定义。更大的窗口大小会略微增加局部上下文生成的成本，但会增加自注意力模块的计算负担。如表6所示，窗口大小为7时，计算成本的增加超过了性能提升，因此，在所有实验中，我们都将 k 设为5。我们使用GFLOPs（即浮点运算次数）来衡量模型的计算复杂度。

k	1	3	5	7	9
$F1$	58.76	60.28	61.58	**62.16**	61.97
十亿次浮点运算	0.46	1.02	1.55	2.21	2.87

表6。局部窗口大小的消融实验 k。同时考虑了F1分数和十亿次浮点运算。

7. 结论

本文正式定义了音频节拍匹配（ABM）任务及其评估协议，旨在帮助

神经网络培养类人编辑技能。为推动后续研究，我们进一步收集了大规模的 *AutoMatch* 数据集，并提出 *BeatX* 作为研究起点。所提出的标签范围有效消除了数据不平衡问题。我们希望这项工作能为计算机视觉研究带来更多启发。

参考文献

[1]Michal Adamkiewicz、Timothy Chen、Adam Caccavale、Rachel Gardner、Preston Culbertson、Jeannette Bohg 和 Mac Schwager。神经辐射世界中的纯视觉机器人导航。*IEEE*《机器人学与自动化快报》，7(2):4606–4613，2022。[2] Omer Bar-Tal、Dolev Ofri-Amar、Rafail Fridman、Yoni Kas- ten 和 Tali Dekel。Text2live：文本驱动的分层图像与视频编辑。载于计算机视觉—*ECCV 2022*：第*17*届欧洲会议，以色列特拉维夫，*2022*年*10*月*23*日至*27*日，会议录，第*XV*部分，第707–723页。施普林格，2022。[3] Suneel Belkhale、Ethan K Gordon、Yuxiao Chen、Siddhartha Srinivasa、Tapomayukh Bhattacharjee 和 Dorsa Sadigh。平衡机器人辅助喂食中的效率与舒适度。载于*2022*年国际机器人与自动化会议（*ICRA*），第4757–4763页。电气电子工程师学会，2022。[4] Srinadh Bhojanapalli、Chulhee Yun、Ankit Singh Rawat、Sashank J. Reddi 和 Sanjiv Kumar。多头注意力模型中的低秩瓶颈，2020。[5] Fatih Cakir、Kun He、Xide Xia、Brian Kulis 和 Stan Sclaroff。用于学习排序的深度度量学习。载于*IEEE/CVF*计算机视觉与模式识别会议会刊，第1861–1870页，2019。[6] Zhe Cao、Tao Qin、Tie-Yan Liu、Ming-Feng Tsai 和 Hang Li。学习排序：从成对方法到列表方法。载于第*24*届国际机器学习大会会刊，第129–136页，2007。[7] Bo Chen、Yichao Wang、Zhirong Liu、Ruiming Tang、Wei Guo、Hongkun Zheng、Weiwei Yao、Muyu Zhang 和 Xi- uqiang He。通过信息共享增强并行深度点击率模型中的显式与隐式特征交互。载于第*30*届*ACM*国际信息与知识管理会议会刊，第3757 3766页，2021 – 。[8] Chi-hau Chen。信号处理手册，第51卷。CRC Press，1988。[9] Haoze Chen 和 Zhijie Zhang。基于新型音频特征的用于车辆类型识别的混合神经网络。科学报告，11(1):7648，2021。[10] Ning Chen、J Stephen Downie、Hai-dong Xiao 和 Yu Zhu。用于翻唱歌曲识别的耳蜗音级类别轮廓。应用声学，99:92–96，2015。[11] Tony G Chen、Becky Miller、Crystal Winston、Stephanie Schneider、Andrew Bylard、Marco Pavone 和 Mark R Cutkosky。Reachbot：一种具有出色伸展能力、适用于崎岖地形的小型机器人。载于*2022*年国际机器人与自动化会议（*ICRA*），第4517–4523页。电气电子工程师学会，2022。[12] Jiankang Deng、Jia Guo、Xiang An、Zheng Zhu 和 Stefanos Zafeiriou。遮挡人脸识别挑战赛：InsightFace赛道报告。载于*IEEE/CVF*国际计算机视觉大会论文集，第1437–1444页，2021年。[13] Nathan Frey、Peggy Chi、Weilong Yang 和 Irfan Essa。自动非线性视频编辑迁移。*CoRR*，abs/2105.06988，2021年。[14] Ohad Fried、Ayush Tewari、Michael Zollh¨ofer、Adam Finkelstein、Eli Shechtman、Dan B Goldman、Kyle Genova、Zeyu Jin、Christian Theobalt 和 Maneesh Agrawala。基于文本的头部特写视频编辑。*ACM*图形学汇刊（*TOG*），38(4):1–14，2019年。[15] Robert Geirhos、J¨orn-Henrik Jacobsen、Claudio Michaelis、Richard Zemel、Wieland Brendel、Matthias Bethge 和 Felix A Wichmann。深度神经网络中的捷径学习。《自然·机器智能》，2(11):665–673，2020年。[16] Huifeng Guo、Ruiming Tang、Yunming Ye、Zhenguo Li 和 Xiuqiang He。DeepFM：一种用于点击率预测的基于因子分解机的神经网络。*arXiv*预印本 *arXiv:1703 04247*，2017年。。[17] 何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习。载于*IEEE*计算机视觉与模式识别会议论文集，第770–778页，2016年。[18] Dexiang Hong、Congcong Li、Longyin Wen、Xinyao Wang 和 Libo Zhang。CVPR 2021通用事件边界检测挑战赛技术报告：级联时序注意力网络（CASTANET）。*arXiv*预印本 *arXiv:2107.00239*，2021年。[19] Dexiang Hong、Xiaoqi Ma、Xinyao Wang、Congcong Li、Yufei Wang 和 Longyin Wen。SC-Transformer++：用于通用事件边界检测的结构化上下文Transformer。*CoRR*，abs/2206.12634，2022年。[20] Hyolim Kang、Jinwoo Kim、Kyungmin Kim、Taehyun Kim 和 Seon Joo Kim。赢得CVPR 2021 Kinetics-GEBD挑战赛：一种对比学习方法。*arXiv*预印本 *arXiv:2106.11549*，2021年。[21] Ankur Kapoor、Ming Li 和 Russell H Taylor。手术辅助机器人的约束控制。载于*ICRA*，第231–236页，2006年。[22] A. Katharopoulos、A. Vyas、N. Pappas 和 F. Fleuret。变换器就是循环神经网络：采用线性注意力的快速自回归变换器。载于国际机器学习大会（*ICML*）论文集，2020年。[23] Minchul Kim、Anil K Jain 和 Xiaoming Liu。AdaFace：用于人脸识别的质量自适应间隔。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第18750–18759页，2022年。[24] Nikita Kitaev、Lukasz Kaiser 和 Anselm Levskaya。Reformer：高效的Transformer。载于国际学习表征会议，2020年。[25] Sharath Koorathota、Patrick Adelman、Kelly Cotton 和 Paul Sajda。像人类一样剪辑：用于自动视频编辑的上下文多模态框架。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第1701–1709页，2021年。

[26]Maria Kyrarini、Fotios Lygerakis、Akilesh Rajavenkatanarayanan、Christos Sevastopoulos、Harish Ram Nambiappan、Kodur Krishna Chaitanya、Ashwin Ramesh Babu、Joanne Mathew 和 Fillia Makedon。医疗保健领域机器人综述。*Technologies*，9(1):8，2021年。[27] Congcong Li、Xinyao Wang、Dexiang Hong、Yufei Wang、Libo Zhang、Tiejian Luo 和 Longyin Wen。用于通用事件边界检测的结构化上下文Transformer。*CoRR*，abs/2206.02985，2022年。[28] Tao Li、Mitsunori Ogihara 和 Qi Li。基于内容的音乐流派分类比较研究。载于第*26*届年度*ACM*国际信息检索研究与发展大会论文集，第282–289页，2003年。[29] Allen Lin、Ziwei Zhu、Jianling Wang 和 James Caverlee。提升对话式推荐系统中的用户个性化。*arXiv*预印本 *arXiv:2302.06656*，2023年。[30] Yang Liu、Fei Han、Fushan Li、Yan Zhao、Maosheng Chen、Zhongwei Xu、Xin Zheng、Hailong Hu、Jianmin Yao、Tailiang Guo 等。采用人工智能认证的喷墨打印不可克隆量子点荧光防伪标签。《自然·通讯》，10(1):2409，2019年。[31] Wantong Lu、Yantao Yu、Yongzhe Chang、Zhen Wang、Chenhui Li 和 Bo Yuan。用于点击率预测的双重输入感知因子分解机。载于第二十九届国际人工智能联合会议国际会议论文集，第3139–3145页，2021年。[32] Marie-Helen Maras 和 Alex Alexandrou。在人工智能时代及深度伪造视频兴起之际，判定视频证据的真实性。《国际证据与证明期刊》，23(3):255–262，2019年。[33] Vivien Marx。大数据面临的重大挑战。《自然》，498(7453):255–260，2013年。[34] Ishaq Azhar Mohammed。人工智能与身份及访问管理之间的交互：一项实证研究。《国际创意研究思想期刊》（*IJCRT*），*ISSN*，2320(2882):668–671，2021年。[35] Anthony Ortiz、Dhaval Negandhi、Sagar R Mysorekar、Shivaprakash K Nagaraju、Joseph Kiesecker、Caleb Robinson、Aditi Khurana、Felipe Oviedo 和 Juan M. Lavista Ferres。用于印度太阳能地点的人工智能数据集。*arXiv*预印本 *arXiv:2202.01340*，2022年。[36] Felipe Oviedo、Srinivas Vinnakota、Eugene Seleznev、Hemant Malhotra、Saqib Shaikh 和 Juan Lavista Ferres。Banknote-Net：用于辅助货币识别的开放数据集。*https://arxiv.org/pdf/2204.03738.pdf*，2022年。[37] Alejandro Pardo、Fabian Caba Heilbron、Juan Le´on Alc´azar、Ali Thabet 和 Bernard Ghanem。MovieCuts：用于剪辑类型识别的新数据集和基准。载于计算机视觉——*ECCV 2022*：第*17*届欧洲会议，特拉维夫，以色列，*2022*年*10*月*23*日至*27*日，论文集，第*VII*部分，第668–685页。施普林格，2022年。[38] Haibo Qiu、Baosheng Yu、Dihong Gong、Zhifeng Li、Wei Liu 和 Dacheng Tao。SynFace：利用合成数据进行人脸识别。载于*IEEE/CVF*国际计算机视觉大会论文集，第10880–10890页，2021年。[39] Ayush K Rai、Tarun Krishna、Julia Dietlmeier、Kevin McGuinness、Alan F Smeaton 和 Noel E O'Connor。《辨识长篇野外视频中的通用事件边界》。*arXiv* 预印本 *arXiv:2106.10090*，2021年。[40] Piyapat Saranrittichai、Chaithanya Kumar Mummadi、Claudia Blaiotta、Mauricio Munoz 和 Volker Fischer。《通过从源域泛化基本视觉因素，克服目标域中的捷径学习》。收录于计算机视觉—*ECCV 2022*：第*17*届欧洲会议，特拉维夫，以色列，*2022*年*10*月*23*日至*27*日，论文集，第二十五部分，第294–309页。施普林格，2022年。[41] Garima Sharma、Kartikeyan Umapathy 和 Sridhar Krishnan。《音频信号特征提取方法的发展趋势》。应用声学，158：107020，2020年。[42] Noam Shazeer。《GLU变体改进Transformer》，2020年。[43]Weichen Shen。《DeepCTR：易用、模块化且可扩展的基于深度学习的点击率模型软件包》。https://github.com/shenweichen/deepctr，2017年。[44] Yaojie Shen、Libo Zhang、Kai Xu 和 Xiaojie Jin。《Auto-transition：学习推荐视频转场效果》。收录于 Shai Avidan、Gabriel J. Brostow、Moustapha Ciss´e、Giovanni Maria Farinella 和 Tal Hassner（编），计算机视觉—*ECCV 2022*—第*17*届欧洲会议，特拉维夫，以色列，*2022*年*10*月*23*日至*27*日，论文集，第三十八部分，《计算机科学讲义》第13698卷，第285–300页。施普林格，2022年。[45] Zhuoran Shen、Mingyuan Zhang、Haiyu Zhao、Shuai Yi 和 Hongsheng Li。《高效注意力：具有线性复杂度的注意力机制》。*CoRR*，abs/1812.01243，2018年。[46] Mike Zheng Shou、Stan Weixian Lei、Weiyao Wang、Deepti Ghadiyaram 和 Matt Feiszli。《通用事件边界检测：事件分割基准》。收录于*IEEE/CVF*国际计算机视觉会议论文集，第8075–8084页，2021年。[47] Jianlin Su。《RoFormer：带旋转位置嵌入的Transformer—zhuiyiai》。技术报告，2021年。[48] Mariarosaria Taddeo 和 Luciano Floridi。《人工智能如何成为造福社会的力量》。《科学》，361(6404)：751–752，2018年。[49] Hongyan Tang、Junning Liu、Ming Zhao 和 Xudong Gong。《渐进式分层提取（PLE）：一种用于个性化推荐的新型多任务学习（MTL）模型》。收录于第*14*届*ACM*推荐系统会议论文集，第269–278页，2020年。[50] Nenad Toma˘sev、Julien Cornebise、Frank Hutter、Shakir Mohamed、Angela Picciariello、Bec Connelly、Danielle CM Belgrave、Daphne Ezer、Fanny Cachat van der Haert、Frank Mugisha等。《惠及社会的人工智能：释放带来积极影响的机遇》。《自然·通讯》，11：2468，2020年，1( )：。[51] Sinong Wang、Belinda Z. Li、Madian Khabsa、Han Fang 和 Hao Ma。《Linformer：具有线性复杂度的自注意力》，2020年。[52] Fen Xia、Tie-Yan Liu、Jue Wang、Wensheng Zhang 和 Hang Li。《学习排序的列表式方法：理论与算法》。收录于第*25*届国际机器学习会议论文集，第1192–1199页，2008年。

[53] Min Xu、Ling-Yu Duan、Jianfei Cai、Liang-Tien Chia、Changsheng Xu 和 Qi Tian。《基于HMM的音频关键词生成》。收录于多媒体信息处理进展—*PCM 2004*：第五届太平洋地区多媒体会议，日本东京，*2004*年*11*月*30*日至*12*月*3*日。论文集，第三部分 *5*，第566–574页。施普林格，2005年。[54] Kimitoshi Yamazaki、Ryohei Ueda、Shunichi Nozawa、Mitsuharu Kojima、Kei Okada、Kiyoshi Matsumoto、Masaru Ishikawa、Isao Shimoyama 和 Masayuki Inaba。《面向老龄化社会的家庭助理机器人》。*IEEE*会刊，100(8)：2429–2441，2012年。[55] Xiaoling Yang、Baohua Tan、Jiehua Ding、Jinye Zhang 和 Jiaoli Gong。《语音活动检测算法的比较研究》。收录于*2010*年国际电气与控制工程会议，第599–602页，2010年。[56] Ajmal Zemmar、Andres M Lozano 和 Bradley J Nelson。《COVID-19疫情期间手术环境中机器人的兴起》。《自然·机器智能》，2(10)：566–572，2020年。[57] Maheen Zulfiqar、Fatima Syed、Muhammad Jaleed Khan 和 Khurram Khurshid。《用于生物特征认证的深度人脸识别》。收录于*2019*年电气、通信与计算机工程国际会议（*ICECCE*），第1–6页。电气电子工程师学会，2019年。