arXiv:2302.05940v1 [cs.SD] 12 Feb 2023

SEMANTICAC：语义辅助音频分类框架

肖一成[1*†]、马越[1*]、李舒妍[1]，、*HantaoZhou*[1]，、*RanLiao*[1]，、李秀[1‡]

[1]清华大学深圳国际研究生院，清华大学，中国

摘要

本文提出 SemanticAC，这是一种语义辅助的音频分类框架，旨在更好地利用语义信息。与将类别标签视为离散向量的传统音频分类方法不同，我们采用语言模型从标签中提取丰富的语义，并优化音频信号与其标签之间的语义一致性。我们验证了，标签中的简单文本信息与先进的预训练模型相结合，能够提供更丰富的语义监督，从而提升性能。具体而言，我们设计了一个文本编码器，用于捕捉标签文本扩展中的语义信息。随后，我们通过音频编码器和相似度计算模块，将音频信号映射至与相应类别标签的语义对齐，从而确保语义一致性。在 ESC-50 和 US8K 两个音频数据集上进行的大量实验表明，我们提出的方法始终优于所比较的音频分类方法。

索引词— 音频、分类、语义

1. 引言

音频分类是音频深度学习和信号处理领域最重要的研究课题之一。这类研究可应用于许多实际领域，包括智能交通 [1], 国家安全 [2]和医疗保健 [3]。音频分类旨在为音频信号分配标签，为一系列任务奠定基础，例如自动语音识别 [4], 关键词检测 [5], 音乐流派识别 [6],等。

在过去十年中，大多数音频分类研究都强调了深度学习的重要性。随着卷积神经网络的快速发展，许多优秀工作将 CNN [7–9] 应用于音频事件分类。它们利用 CNN 从音频信号中提取特征，并设计出多种损失函数，以获得具有判别性的特征。受 Transformer 强大编码能力的启发，许多方法 [10]，[11–13] 被用于对音频信号进行建模，并取得了出色的性能。其中，AST [11]是首个在音频分类中引入自注意力

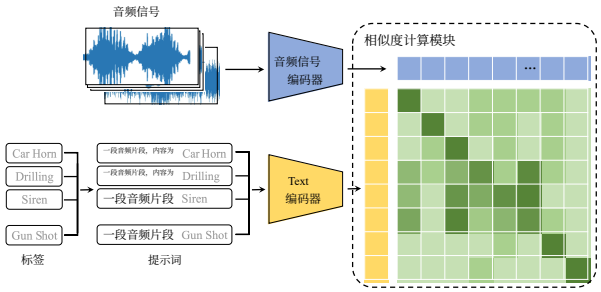

图1。我们从标签中提取语义，并优化音频信号与其对应标签之间的语义一致性。

机制的模型。HTS-AT [12]设计了层次化的 Transformer 结构，并取得了巨大成功。然而，仅使用单一信号源无法充分利用多媒体中包含的丰富信息。近年来，多模态方法 [14–17] 日益受到关注。例如，MBT [14]通过视听融合获得了更具判别性的音频特征表示。AudioCLIP [15] 成功将音频模态整合到 CLIP [18], 中，证明了通过视觉和自然语言监督学习音频表示是有效的。然而，大多数音频数据集缺乏视觉信号，这促使我们探索仅利用自然语言高效辅助音频信号建模的潜力。

本文提出了一种简单而有效的音频分类框架，即 SemanticAC。如图1所示，其基本思路是从分类标签中提取语义信息，并利用这些信息辅助音频建模。具体而言，我们设计了一个文本编码器，将用于标签文本扩展的提示词映射为语义表示。为了对齐基于 Transformer 的结构所获得的音频特征与标签语义，我们设计了一个轻量且高效的相似度计算模块，并尝试通过对比学习缩小这两种模态之间的差距。

我们在两个数据集 ESC-50 [19]和 US8K [20]上进行了大量实验，以验证所提出的 SemanticAC 的有效性。本文的贡献可概括如下：

- W我们提出了一种有效的语义辅助音频 -文本

*共同贡献。 †工作是在清华大学实习期间完成的——清华大学国际研究生院。 ‡通讯作者。

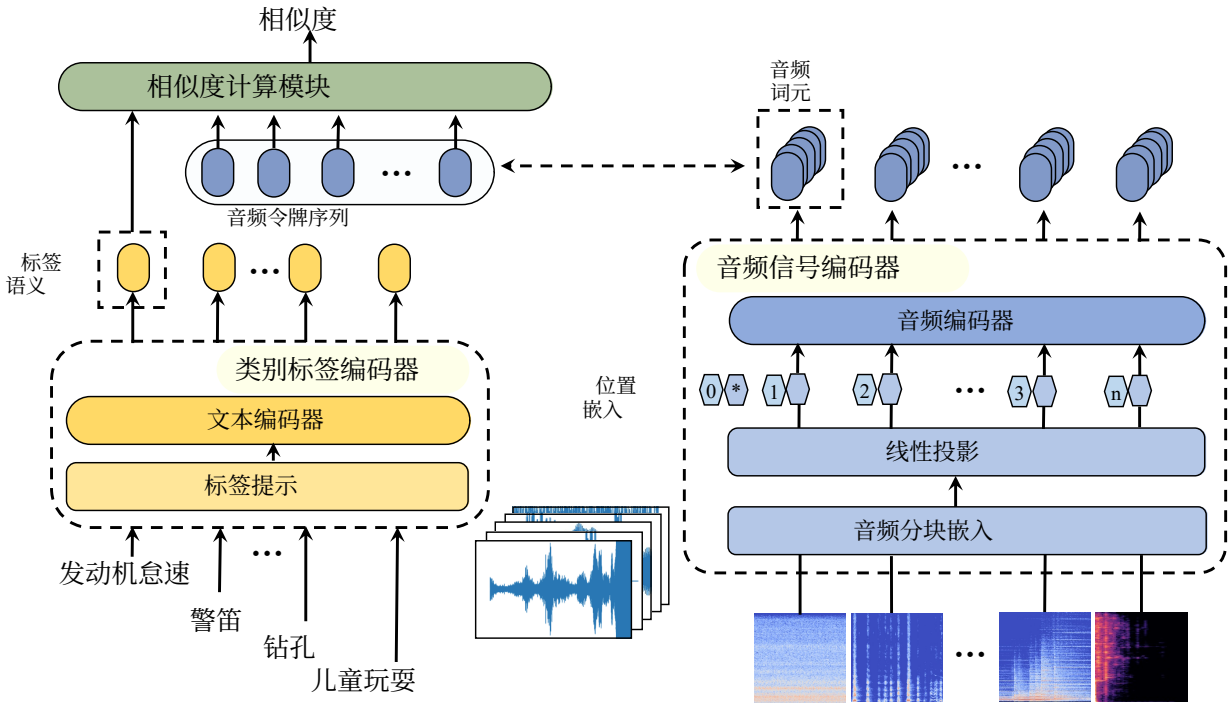

图 2。我们提出的语义辅助音频分类框架概览。

模态框架，名为 SemanticAC，用于音频分类。与传统方法将标签视为离散向量不同，我们的方法充分利用了标签中的语义信息。

- 我们开发了一种轻量且高效的相似度计算模块，该模块完全基于卷积神经网络结构，称为 CSCM，用于将音频特征与语义对齐。
- 我们在两个数据集 ESC-50 [19] 和 US8K [20] 上取得了稳定且显著的改进。

我们将在以下章节中详细介绍我们的框架和实验验证。

2. SEMANTICAC

给定一批音频信号，我们以特定采样率将其转换为单声道。随后将其转换为梅尔频谱图，记为 $\{x_i\} \in R^{F\times T}$，其中$F$，$T$ 和 i 分别表示频谱特征的维度、时间帧的数量和音频样本的索引。我们的目标是将每个音频样本 $\{x_i\}$ 映射到其类别标签 $\{y_i\} \in Y$：

$$F\left(x_i|\theta\right) : R^{F\times T} \longmapsto Y \tag{1}$$

其中 θ 是映射模型的参数。

我们不再将类别标签视为离散向量（即独热向量），而是通过 SemanticAC 从标签中提取丰富的语义，并利用这些语义信息辅助音频表征学习。SemanticAC 的概览如图 2 所示。我们设计了一个文本编码器，从类别标签中提取语义 $\{T_i\} \in \Theta^C$，并设计了一个音频编码器来提取音频令牌。随后，我们开发了一个相似度计算模块，以获得音频特征 $\{A_i\} \in \Theta^C$，并将其与 Θ^C中的标签语义对齐。其中，Θ^C 表示维度为 C 的投影共享嵌入空间，使音频信号能够正确映射到对应的类别标签。

2.1. 文本-音频多模态编码器

在本小节中，我们将介绍如何提取类别标签的语义和音频特征令牌。我们以 CLIP [18], 为基础设计骨干网络，其中包括一个类别标签编码器和一个音频信号编码器。

类别标签编码器。该编码器用于从标签中提取语义。我们没有将离散类别标签直接输入编码器，而是设计了提示词“an audio clip of $[LABEL]$”，其中 $LABEL$ 表示相应音频信号的标签，用于生成文本扩展作为输入。这基于我们的实证发现：简单的短语描述有助于从标签中提取语义信息，具体内容将在第3.4节中详细介绍。为了从标签文本扩展中提取丰富的语义信息，我们

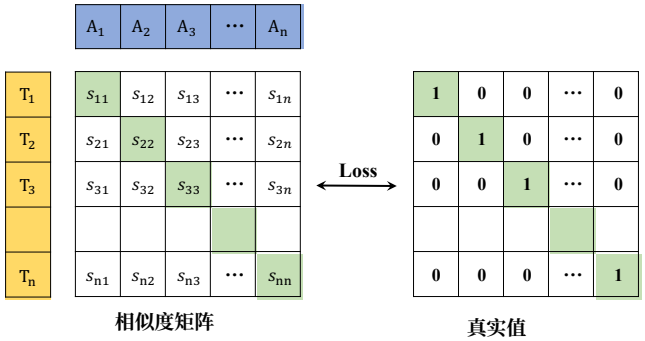

图 3。该图展示了音频和文本特征的相似度矩阵 $\mathbf{S}_{n\times n}$，其中 $s_{ij} = A_i \cdot T_j (i,\ j = 1,\ 2,\ \cdots,\ n)$ 及其对应的真实值。

设计了一个多层基于 Transformer 的文本编码器，其灵感来源于 Transformer 强大的文本建模能力。经过文本编码后，我们得到一个 C 维向量，用于表示标签语义。

音频信号编码器。首先，我们将音频信号的频谱图切分为图像块，以便更好地捕捉不同时间帧之间频率单元的对应关系。随后，我们通过一个线性层将这些图像块投影为一个序列。我们用一个基于 Transformer 的预训练编码器替换 CLIP 的图像头 ViT [21]，该编码器采用遵循 HTS-AT [12] 的窗口注意力分层 Transformer，最终得到一个特征令牌序列。

2.2. 相似度计算模块

我们设计了一个名为 CSCM 的相似度计算模块，用于捕捉标签语义与音频信号之间的相关性。传统方法会将高维音频令牌序列直接投影到嵌入空间，并进行大量逐元素计算。相比之下，我们采用一系列卷积核大小适当的卷积网络来减少逐元素乘法，从而将参数规模控制在较小范围内。为生成更丰富的特征表示，我们采用卷积注意力机制 [22]。具体而言，该机制首先将音频令牌序列重塑为 $3D$ 个特征图 M，位于 $[d \times h \times w]$中，其中d, h 和 w 分别为 M 的深度、高度和宽度，随后输出一个音频特征向量，其维度与嵌入空间中文本表示的维度 C 相同。最后，我们采用对比学习优化整个网络。如公式（2）所示，我们通过余弦相似度衡量两种模态表示之间的距离。

$$s_{ij} = \frac{A_i \cdot T_j}{\|A_i\| \cdot \|T_j\|} = \frac{\sum_{m=1}^{C} A_i^m \times T_j^m}{\sqrt{\sum_{m=1}^{C}(A_i^m)^2} \times \sqrt{\sum_{m=1}^{C}(T_j^m)^2}} \tag{2}$$

得到的相似度矩阵 $\mathbf{S}_{n\times n}$ 如图 3 所示，其中n 为批量大小。$\mathbf{S}_{n\times n}$ 中的对角元素表示正样本对，其余所有元素表示负样本对。我们通过矩阵不同轴上的交叉熵损失 $CE()$，最小化正样本对之间的距离，并最大化负样本对之间的距离。损失函数表示如下，其中 $\mathbf{S}'$ 表示$\mathbf{S}$的转置：

$$Loss = \frac{1}{2}\left[CE\left(\mathbf{S}, Y\right) + CE\left(\mathbf{S}', Y\right)\right], \tag{3}$$

3. 实验

在本节中，我们在两个数据集上评估所提出的方法：ESC-50 [19]和 US8K [20]。

3.1. 数据集

ESC-50 [19]数据集包含 2000 个短音频片段，涵盖 50 类不同的环境声音事件，并分为 5 折。US8K [20]是一个音频数据集，包含 8732 段城市声音的标注音频片段，分为 10 类和 10 折。在数据处理过程中，我们将 ESC-50 的信号重采样至 44100Hz 用于训练，并重采样至 32000Hz 用于评估。对于 US8K 中的样本，我们均将其重采样至 44100Hz 用于训练和评估。

3.2. 实现细节

音频编码器已在 AudioSet [27]上进行预训练。为了微调我们的语义辅助框架，我们使用 SGD [28]，采用权重衰减 $5e-4$ 和批量大小 16 来优化参数。我们使用 ExponentialLR 策略来调度学习率，初始值为 $8e-5$，衰减因子 γ 设置为 0。96。输入文本使用词汇表中大小为 49152 的小写字节对进行编码，与 CLIP 相同。第 2.2 节中提到的 M 的维度 d、h 和 w 分别设置为 768、8 和 8。标签语义和音频特征向量维度 C 设置为 1024。考虑到计算资源，我们将最大序列长度设置为 76。按照 [15],，我们应用了多种

表 1。ESC-50 上的评估。

模型	预训练	TOP-1 准确率	平均值（准确率 %）
SepTr [13]	-	91.13	-
EAT-M [23]	✓	96.3	96.3
AST [11]	✓	-	95.6
HTS-AT [12]	✓	97	97
XDC [24]	✓	85.4	-
CrissCross [25]	✓	90.5	-
AudioCLIP [15]	✓	-	97.15
AVID [26]	✓	89.2	-
SemanticAC*(我们的方法)	✓	96.5	96.5
SemanticAC(我们的方法)		**97.25**	**97.25**

表2。US8K 上的音频分类准确率

模型	额外训练 Data	准确率（%）
1DCNN [7]	✓	89
DenseNet [9]		87.42
ESResNeXt [8]		89.14
AudioCLIP [15]	✓	90.07
EAT-M [23]	✓	90
SemanticAC（我们的方法）	✓	**91.34**

数据增强方法，例如随机裁剪、随机噪声等，以避免过拟合。

3.3. 与最先进水平的比较

如表 1 所示，我们在 ESC-50 数据集上取得了最先进水平（SOTA），准确率达到 97.25%。仅使用音频模态时，SemanticAC 的表现比 HTS-AT 高 0.25%，这表明了我们框架的有效性。随后，我们采用与 ESC-50 相同的训练策略，在 US8K 上进行评估。表 2 中的实验结果显示，我们的性能为 91.34%，比 AudioCLIP [15] 高 1.27%，比 EAT-M [23] 高 1.34%。图 4 的可视化结果展示了我们的方法在 US8K 的第 8th 折上的表现。

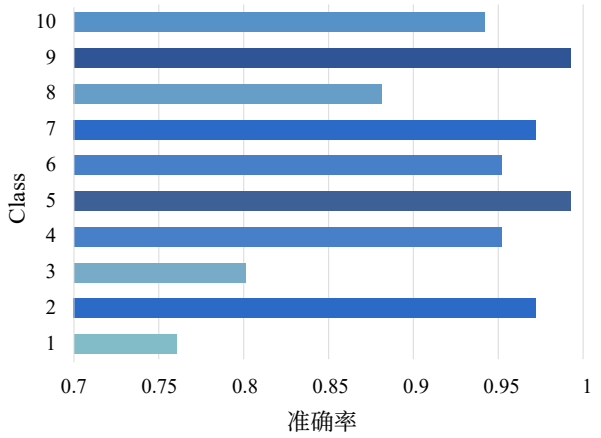

图 4。US8K 第 8th 折中 10 个类别的准确率。类别编号 $1-10$ 分别代表以下类别：空调、汽车喇叭、儿童玩耍、狗吠、钻孔、发动机怠速、枪声、手提钻、警笛、街头音乐。

3.4. 消融研究

我们开展了两组实验，以验证本方法各个组成部分的有效性。

文本辅助与 CSCM。我们的相似度计算模块通过结合注意力机制和 CNN模块，将音频特征与标签的语义表示对齐，并最终计算余弦相似度。

表3。 4th ESC-50折上不同配置的结果。

	文本辅助	Sim. Cal. 模块	ACC. %
基线	-	-	94.2
SemanticAC*	✓	-	96.5
SemanticAC	✓	SeqLSTM	97.25
SemanticAC	✓	SeqTransf	97.5
SemanticAC	✓	CSCM（本文方法）	**98**

表4. ESC-50 上不同提示词作为文本输入方式的准确率比较。

提示词	平均值（准确率，%）
$[LABEL]$	97.15(±0.17)
$a\ clip\ of\ [LABEL]$	97.22(±0.4)
$an\ audio\ clip\ of\ [LABEL]$	97.25(±0.25)

我们采用若干卷积层，卷积核大小为 3×3 和 1×1，以实现特征对齐。如表 3 所示，SemanticAC* 代表仅采用文本辅助的方法，在 ESC-50 的第 4th 折上准确率达到 96.5%，比基线高 2.3%；使用我们的相似度计算模块后，准确率还可提升 1.5%。我们在 ESC-50 的第 4th 折上对该模块进行了验证，并将其与采用 Transformer结构的 SeqTransf 以及采用 LSTM结构的 SeqLSTM 进行比较，这两种方法均见于 CLIP4CLIP [29] 。如表 3 所示，CSCM 能够实现高效的融合与对齐。我们认为，文本辅助能够提供丰富的语义监督，从而提升音频建模的效果。此外，CSCM 有助于缩小文本与音频之间的差距。

标签提示词。我们在 ESC-50 上进行了实验，结果如表 4 所示。与直接使用“$[LABEL]$”作为输入相比，用于标签文本扩展的提示词“a clip of $[LABEL]$”可将准确率提高 0.07%。与此同时，通过进一步丰富文本输入，例如“an audio clip of $[LABEL]$”，有助于类别标签编码器从标签中提取丰富的重要语义。它利用固定搭配扩展文本，从而增强文本建模。

4. 结论

本文提出了一种音频分类框架 SemanticAC，该框架利用类别标签中的语义信息来辅助音频表征学习。我们在两个音频数据集 ESC-50 和 US8K 上取得了显著提升。未来，我们将探索利用更多信号源来辅助音频分类。

5. 参考文献

[1] D. Williams、D. De Martini、M. Gadd、L. Marchegiani 和 P. Newman，“Keep off the grass: Permissible driv- ing routes from radar with weak audio supervision”，载于*ITSC*。IEEE，2020年，第 1–6 页。

[2] S. Shan、J. Liu 和 Y. Dun，“基于深度学习的声纹识别前景”，载于 *JPCS*。IOPPublishing，2021年，第1848卷，第012046页。

[3] M. Esposito、G. Uehara 和 A. Spanias，“面向医疗保健应用的音频分类量子机器学习”，载于*IISA*。IEEE，2022年，第1–4页。[4] C. Zoril˘a 和 R. Doddipatla，“利用目标声源提取进行说话人增强，以实现鲁棒的自动语音识别”，载于*ICASSP*。IEEE，2022年，第6297–6301页。[5] G. Chen、C. Parada 和 G. Heigold，“使用深度神经网络的小型化关键词检测”，载于*ICASSP*。IEEE，2014年，第4087–4091页。[6] D. Ghosal 和 M. H. Kolekar，“使用深度神经网络和迁移学习的音乐流派识别”，载于*Interspeech*，2018年，第2087–2091页。[7] S. Abdoli、P. Cardinal 和 A. L. Koerich，“使用一维卷积神经网络进行端到端环境声音分类”，*ESWA*，第136卷，第252–263页，2019年。[8] A. Guzhov、F. Raue、J. Hees 和 A. Dengel，“Esresne (x) t-fbsp：学习鲁棒的音频时频变换”，载于*IJCNN*。IEEE，2021年，第1–8页。[9] K. Palanisamy、D. Singhania 和 A. Yao，“重新思考用于音频分类的卷积神经网络模型”，*arXivpreprintarXiv:2007.11154*，2020年。

[10]A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，《注意力即所需的一切》，*NeurIPS*，第 30 卷，2017年。

[11] Y. Gong、Y.-A. Chung 和 J. Glass，《AST：音频频谱图 Transformer》，*arXiv* 预印本 *arXiv:2104.01778*，2021年。

[12] K. Chen、X. Du、B. Zhu、Z. Ma、T. Berg-Kirkpatrick 和 S. Dubnov，《HTS-AT：用于声音分类和检测的分层令牌语义音频变换器》，载于*ICASSP*。IEEE，2022年，第 646–650 页。

[13] N.-C. Ristea、R. T. Ionescu 和F.S.Khan，《SepTr：用于音频频谱图处理的可分离 Transformer》，*arXiv* 预印本 *arXiv:2203.09581*，2022年。

[14] A. Nagrani、S. Yang、A. Arnab、A. Jansen、C. Schmid 和 C. Sun，“多模态融合的注意力瓶颈”，*NeurIPS*，第 34 卷，第 14200–14213 页，2021年。

[15] A. Guzhov、F. Raue、J. Hees 和 A. Dengel，《AudioCLIP：将 CLIP 扩展至图像、文本和音频》，载于*ICASSP*。IEEE，2022年，第 976–980 页。[16] Y. Ma、T. Yang、Y. Shan 和 X. Li，《SimVTP：采用掩码自编码器进行简单的视频文本预训练》，*arXiv* 预印本 *arXiv:2212.03490*，2022年。[17] Y. Ma、Y. Wang、Y. Wu、Z. Lyu、S. Chen、X. Li 和 Y. Qiao，《用于视频中人类动作推理的视觉知识图谱》，载于第 *30* 届 *ACM* 国际多媒体会议论文集，2022年，第 4132–4141 页。[18] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，《从自然语言监督中学习可迁移视觉模型》，载于*ICML*。PMLR，2021年，第 8748–8763 页。[19] K. J. Piczak，《ESC：环境声音分类数据集》，载于*ACM MM*，2015年，第 1015–1018 页。[20] J. Salamon、C. Jacoby 和 J. P. Bello，《面向城市声音研究的数据集与分类体系》，载于*ACMMM*，2014年，第 1041–1044 页。[21]A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner 等，《一张图像胜过 16×16 个词：用于大规模图像识别的变换器》，*arXiv* 预印本 *arXiv:2010.11929*，2020年。[22] S. Woo、J. Park、J.-Y. Lee 和 I. S. Kweon，《CBAM：卷积块注意力模块》，载于*ECCV*，2018年，第 3–19 页。[23] A. Gazneli、G. Zimerman、T. Ridnik、G. Sharir 和 A. Noy，《端到端音频强势回归：通过增强提升高效音频分类网络》，*arXiv* 预印本 *arXiv:2204.11479*，2022年。[24]H. Alwassel、D. Mahajan、B. Korbar、L. Torresani、B. Ghanem 和 D. Tran，《通过跨模态音视频聚类进行自监督学习》，*NeurIPS*，第 33 卷，第 9758–9770 页，2020年。[25] P. Sarkar 和 A. Etemad，《通过放宽跨模态时间同步性进行自监督视听表征学习》，*arXiv* 预印本 *arXiv:2111.05329*，2021年。[26] P. Morgado、N. Vasconcelos 和 I. Misra，《通过跨模态一致性进行视听实例判别》，载于*CVPR*，2021年，第 12475–12486 页。[27] J. F. Gemmeke、D. P. Ellis、D. Freedman 等，《AudioSet：音频事件本体与人工标注数据集》，载于*ICASSP*。IEEE，2017年，第 776–780 页。

[28] B. T. Polyak 和 A. B. Juditsky，《通过平均加速随机逼近》，*SIAM J CON-TROL OPTIM*，第 30 卷，第 4 期，第 838–855 页，1992年。

[29] H. Luo、L. Ji、M. Zhong、Y. Chen、W. Lei、N. Duan 和 T. Li，“Clip4clip：关于 CLIP 用于端到端视频片段检索和字幕生成的实证研究”，*Neurocomput-ing*，第 508 卷，第 293–304 页，2022年。