arXiv:2302.09328v1 [cs.MM] 18 Feb 2023

SSVMR：基于显著性的视频-音乐检索自训练

Xuxin Cheng、*Zhihong Zhu*、*Hongxiang Li*、*Yaowei Li*、*Yuexian Zou**

ADSPLAB，电子与计算机工程学院，北京大学{chengxx、zhihongzhu、lihongxiang、ywl}@stu.pku.edu.cn，zouyx@pku.edu.cn

摘要

随着短视频的兴起，为视频选择合适的背景音乐（BGM）的需求显著增加，视频-音乐检索（VMR）任务逐渐受到研究界的广泛关注。与其他跨模态学习任务一样，现有的 VMR 方法通常试图在特征空间中衡量视频与音乐之间的相似度。然而，这些方法：（1）忽视了不可避免的标签噪声；（2）未能增强捕捉关键视频片段的能力。本文提出了一种新颖的基于显著性的自训练框架，称为 SSVMR。具体而言，我们首先采用半监督方法，充分利用训练数据集所包含的信息，以抑制标签噪声问题的不利影响，其中采用了自训练方法。此外，我们通过在片段级混合两个视频，同时保留两个原始视频的局部性，提出了捕捉视频显著性的方法。受自然语言处理中的反向翻译启发，我们还进行了反向检索，以获取更多训练数据。在 MVD 数据集上的实验结果表明，与之前的最佳模型相比，我们的 SSVMR 在 R@1 指标上取得了 34.8% 的相对提升，实现了大幅领先的当前最优性能。

索引词— 视频-音乐检索、跨模态匹配、自训练、标签噪声

1. 引言

背景音乐 （BGM）可以调节氛围，增强视频的情感表达，并为观众带来沉浸感。随着短视频的兴起，为视频选择合适背景音乐的需求显著增加。然而，手动完成这项工作需要大量专业知识。为解决这一问题，视频-音乐检索 （VMR）日益受到研究人员的关注。视频-音乐检索旨在从音频库中为输入视频匹配合适的背景音乐。

近年来，现有的视频-音乐检索方法尝试在特征空间中衡量视频与音乐之间的相似度。与其他跨模态任务一样， [1, 2, 3], 一些传统的视频-音乐检索方法通过利用外部数据进行跨模态匹配，例如元数据 [4, 5, 6]。元数据通常以快速且简单的方式提供有关视频或音乐内容的关键信息，从而使多媒体内容像文本一样易于检索。 [4]使用专辑封面的颜色直方图表示来索引音乐片段。 [5]通过衡量视频情感标签与音乐情感标签的相似度，为视频检索背景音乐。 [6] 将情感标签作为联合约束，对音乐和视频特征提取器进行预训练，从而更好地对齐视频与音乐模态的嵌入。不幸的是，元数据并不总是可用，尤其是在数据规模较大时。近年来，部分研究尝试采用不依赖元数据的基于内容的方法来完成视频-音乐检索任务。 [7] 提出了一种双分支深度网络，在考虑模态间和模态内关系的情况下关联视频与音乐。 [8] 提出了一种跨生成策略，也能更好地对齐视频和音乐的潜在嵌入。

尽管现有视频-音乐检索模型已经取得了令人鼓舞的进展，但我们发现它们存在两个主要问题：**（1）忽视视频-音乐检索中不可避免的标签噪声 [9]**。标签噪声很容易导致模型过拟合 [10, 11]。与其他跨模态任务相比，视频-音乐检索任务可能包含更多标签噪声，原因有二：a）视频-音乐检索数据集的标签具有主观性，不同专家可能会为同一视频选择截然不同的背景音乐；b）视频-音乐检索数据集 HIMV-50K [12]和 MVD [13] 均由三种视频-音乐 对组成，包括官方音乐视频、恶搞音乐视频和用户生成视频，从而导致数据与标签之间的关系存在一定差异。

（2）忽视了增强捕捉关键视频片段的能力。直观而言，视频中的不同片段在确定对应音乐方面发挥着不同的作用。视频中可能存在一小段内容，发挥着最为重要的作用。然而，以往的工作将每个视频片段一视同仁，这可能导致模型性能下降。

在本文中，我们提出了基于**S**显著性**S**的**V**视频-**M**音乐**R**检索（**SSVMR**）自训练方法，以解决上述两个问题。针对第一个问题，我们使用高斯混合模型（GMM）筛选出更有可能错误的样本，并通过锐化函数将其原始标签替换为软标签。随后，我们引入 R-Drop [14] 损失，通过约束同一样本输入两次后得到的两个预测分布，来减小自训练中的确认偏误。

*通讯作者

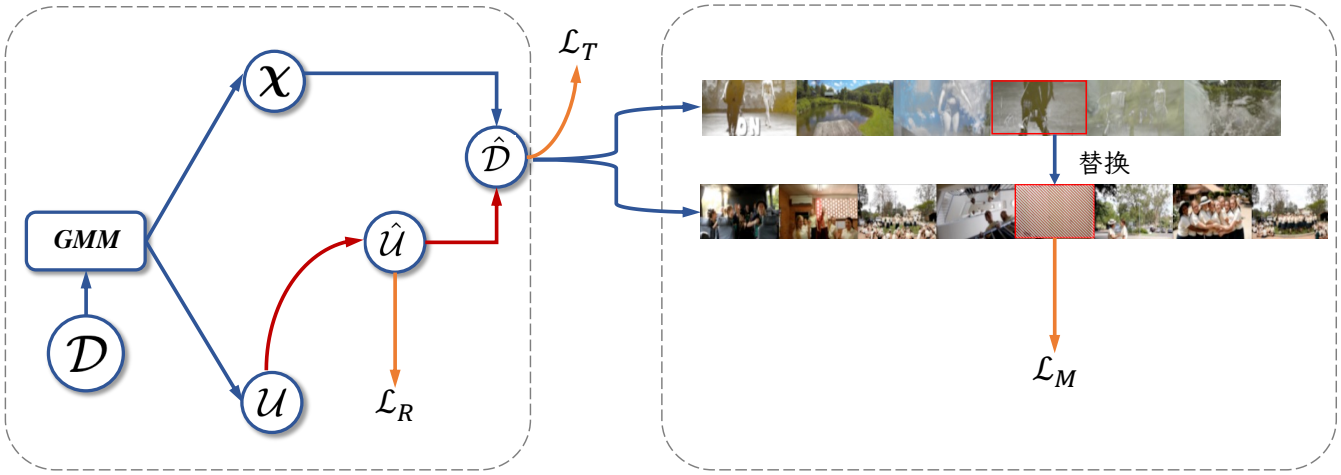


图 1： 我们提出的框架架构，由两个部分组成：(1) 使用 R-Drop 的半监督自训练；(2) 依靠显著性的片段级混合。 $\mathcal{D}$ 表示原始的视频-音乐 对，包括有标签集 $\mathcal{X}$ 和 ^ ^ ^ 无标签集 $\mathcal{U}$。 $\mathcal{U}$ 表示带有新软标签的 $\mathcal{U}$。 $\mathcal{D}$ 表示 $\mathcal{X}$ 和 $\mathcal{U}$的组合。

针对第二个问题，我们利用显著性在片段级混合两个视频。这样可以保留与预测相关的更多重要信息，同时保留两个原始视频的局部性。此外，受 NLP 中回译 [15] 的启发，我们从音乐-视频检索模型执行反向检索，以扩充数据集。随后 [12], 我们在 HIMV- 50K [12]数据集上训练模型，并在 MVD [13] 数据集上进行评估。实验和进一步分析表明，SSVMR 能够有效处理标签噪声，并提升模型捕捉关键视频片段的能力，其性能大幅超越此前的最先进方法。据我们所知，我们的框架是首个处理标签噪声并在 VMR 任务中考虑显著性的方法。

2. 方法

本节首先在第 2.1 节介绍视频-音乐检索（VMR）的问题定义，并在第 2.2 节介绍骨干网络，随后详细描述我们的框架。如图 1 所示，该框架主要由第 2.3 节介绍的带有 R-Drop 模块的自训练、第 2.4 节介绍的依靠显著性模块的片段级混合，以及第 2.5 节介绍的反向检索模块组成。

2.1. 问题定义

用于视频-音乐检索任务的数据集通常包含视频-音乐对，可表示为 $\mathcal{D} = \{(v,\ m)\}$，其中 v 是视频的嵌入特征，m 是对应背景音乐的嵌入特征。对于视频集合 $\mathcal{V}$ 和音乐集合 $\mathcal{M}$，视频-音乐检索方法旨在检索音乐候选项列表 $\mathcal{C}(m) \subset \mathcal{M}$，其中 m 中的每首音乐都是来自 $\mathcal{V}$的查询视频 v 的潜在匹配项。

2.2. 骨干网络

我们使用 CBVMR [7]作为骨干网络，该网络采用双分支神经网络，并结合模态间排序约束与软性模态内结构约束，构建统一嵌入网络。它通过应用三元组损失 $\mathcal{L}_T$ 来最小化经验风险，具体如下：

$$\begin{aligned}\mathcal{L}_T = \sum_i \ell_i \ &= \lambda_1 \sum_{i\neq j} \max\left(0, v_i^\top m_j - v_i^\top m_i + e\right) \\ &+ \lambda_2 \sum_{i\neq j} \max\left(0, m_i^\top v_j - m_i^\top v_i + e\right) \\ &+ \lambda_3 \sum_{i\neq j\neq k} C_{ijk}(v)\left(v_i^\top v_j - v_i^\top v_k\right) \\ &+ \lambda_4 \sum_{i\neq j\neq k} C_{ijk}(m)\left(m_i^\top m_j - m_i^\top m_k\right)\end{aligned} \tag{1}$$

其中，λ_1 和 λ_2 平衡视频到音乐匹配和音乐到视频匹配中的跨模态排序损失的影响。每种模态内部的软模态内结构损失的影响则分别由 λ_3 和 λ_4进行平衡。 e 是三元组损失的边界，函数 $C(\cdot)$定义如下：

$$C_{ijk}(x) = \text{sign}\left(x_i^\top x_k - x_i^\top x_j\right) - \text{sign}\left(\widetilde{x}_i^\top \widetilde{x}_k - \widetilde{x}_i^\top \widetilde{x}_j\right) \tag{2}$$

其中，x_i、x_j和 x_k 是多模态空间中的可训练特征，而 $\widetilde{x}_i$、$\widetilde{x}_j$和 $\widetilde{x}_k$ 是通过嵌入网络之前的模态内特征。我们首先使用 $\mathcal{L}_T$ 对模型进行预热，使其能够执行初步的视频-音乐检索任务，同时避免对噪声标签过拟合，然后在剩余的训练轮次中执行SSVMR。

2.3. 使用R-Drop的半监督自训练

如 [16], 所述，噪声样本通常在早期阶段具有较高的损失。在训练过程中，损失

干净样本和噪声样本的分布往往服从两个高斯分布 [17], , 其中噪声样本的损失具有更大的均值。因此，我们采用广泛使用的 GMM，根据每个输入样本的损失来区分噪声样本。我们使用包含 2 个成分的 GMM，并通过期望最大化（EM）算法对观测值进行拟合。具体而言，我们首先计算每个样本属于均值较大的高斯成分 g的概率 w_i，即属于噪声样本的概率，如下所示：

$$w_i = p\left(g \mid \ell_i\right) \tag{3}$$

然后，我们为概率设置一个阈值 τ。原始数据集 $\mathcal{D}$ 被划分为两部分，包括有标签集$\mathcal{X}$ 和无标签集 $\mathcal{U}$。无标签集 $\mathcal{U}$ 中样本的标签很可能是错误的，因此我们用新标签替换它。我们通过锐化模型的预测分布来生成新的软标签 $\hat{}m$ ，将无标签集 $\mathcal{U}\hat{}$ 转换为有标签集 $\mathcal{U}$，从而促使模型生成熵更低的预测：

$$\mathcal{X} = \{(v_i, m_i) \mid v_i \in \mathcal{D}, w_i \leq \tau\} \tag{4}$$

$$\mathcal{U} = \{(v_i) \mid v_i \in \mathcal{D}, w_i > \tau\} \tag{5}$$

$$\hat{m} = \text{Sharpen}(p(v, \theta), T) \tag{6}$$

$$\hat{\mathcal{U}} = \{(v_i, \hat{m}_i) \mid v_i \in \mathcal{U}\} \tag{7}$$

$$\hat{\mathcal{D}} = \mathcal{X} \cup \hat{\mathcal{U}} \tag{8}$$

其中， $p(v,\ \theta)$表示样本 v的预测结果，Sharpen (·)是自训练中常用的温度锐化函数， T 表示温度：

$$\text{Sharpen}(p, T)_i = p_i^{\frac{1}{T}} / \sum_{j=1}^{L} p_j^{\frac{1}{T}} \tag{9}$$

基于丢弃法的模型在训练和推理之间存在不一致性，因为丢弃法在训练时使用，而在推理时不使用；这种不一致性在噪声样本中会更加明显，并对模型造成更大的损害。因此，我们引入了 R-Drop [14] 损失 $\mathcal{L}_R$ 来解决这一不一致性：

$$\begin{aligned}\mathcal{L}_R = \sum_{m \in \mathcal{U}} \frac{1}{2} \left(\mathcal{D}_{KL}\left(p_1(m, \theta) \| p_2(m, \theta)\right)\right. \\ \left. + \mathcal{D}_{KL}\left(p_2(m, \theta) \| p_1(m, \theta)\right)\right)\end{aligned} \tag{10}$$

其中， $p_1(m,\ \theta)$和 $p_2(m,\ \theta)$是通过两次输入同一样本得到的两个预测分布。 $\mathcal{D}_{KL}(a\|b)$表示*Kullback-Leibler* 散度，用于对这两个概率分布进行正则化。我们将带有原始标签的干净样本与带有ˆ生成软标签的噪声样本相结合，得到新的数据集 $\mathcal{D}$。

2.4. 基于显著性的片段级混合

视频中可能存在一小段内容，对检索音乐的结果起着最关键的作用。通过捕捉关键的视频片段，模型可以借助引入预训练技术，更容易地进一步提升性能 [18]。受 [19, 20, 21], 启发以改进我们的框架捕捉关键视频信息的能力，我们在片段级别混合两个视频。首先，我们计算 $\mathcal{L}_T$ 相对于输入嵌入 v 的梯度，并将其幅度用作显著性：

$$s = \|\partial \mathcal{L} / \partial e\|_2 \tag{11}$$

其中， $\|\cdot\|_2$ 表示L2范数。对于两个视频-音乐 配对，分别表示为(v_1, m_1)和(v_2, m_2)，我们从 v_1 中选取显著性最低的片段 x_1^S ，并用 v_2 中显著性最高的片段x_2^S 替换它，从而得到新的视频嵌入ˆv ，该嵌入由$(v_1^L;\ v_2^S;\ v_1^R)$ 拼接而成，其中v_1^L和v_1^R分别是原始视频 x_1中位于 x_1^S左侧和右侧的嵌入。两个视频所选片段的长度相同，计算方式如下：

$$\ell_{v1} = \ell_{v2} = \max\left(\min\left(\left[\lambda_0 \left|v_1\right|\right], \left|v_2\right|\right), 1\right) \tag{12}$$

其中， ℓ_{v1} 和 ℓ_{v2} 分别表示长度， λ_0 表示初始混合增强比例。我们计算增强输出与每个样本原始视频嵌入之间的三元组损失 $\mathcal{L}_T$，并按权重 λ 将其组合，得到混合损失 $\mathcal{L}_M$：

$$\lambda = \left|v_2^S\right| / |\hat{v}| \tag{13}$$

$$\mathcal{L}_M = \lambda \mathcal{L}_T\left(\hat{v}, m_1\right)) + (1 - \lambda)\, \mathcal{L}_T\left(\hat{v}, m_2\right) \tag{14}$$

采用三元组损失和混合损失时，最终训练目标如下：

$$\mathcal{L} = \mathcal{L}_T + \mathcal{L}_R + \mathcal{L}_M \tag{15}$$

2.5. 反向检索

在按照上述方式训练视频音乐检索模型之前，我们进行反向检索以获取更多数据。在神经机器翻译任务中，回译 [15] 被广泛用于扩充平行训练语料库。受此启发，我们首先使用原始数据集训练一个音乐-视频检索模型，然后将每段音乐输入该模型，以获取与其最相似的三个视频。考虑到视频与音乐之间的匹配标准比其他跨模态任务（例如图像到文本检索）更加模糊，我们随机选择其中一个视频，并将其与该音乐组合成一个新的数据对。

3. 实验

3.1. 数据集与设置

随后 [12], 我们在 HIMV-50K [12]数据集上训练模型，并在 MVD [13] 数据集上对其进行评估。HIMV-50K 是一个大型音乐视频集合，对应于 YouTube-8M [22] 数据集中标注为“music video”的部分片段，由 50,000视频-音乐 对组成，且没有标注。需要注意的是，虽然这些片段的音频始终包含音乐，但视频内容可能从专业宣传片到静态图像的业余蒙太奇不等，因此这些片段中视频部分的质量具有很大的异质性。因此，我们使用 MVD [13] 数据集来评估模型。该数据集包含 2,212 个专业质量的音乐视频片段。每个片段的平均时长为 4 分钟。我们从中随机选取了 2000 个片段来评估模型。

在本文中，参数 λ_1、λ_2、λ_3、λ_4 和 e 分别设为 3、1、0.2、0.2 和 6。温度 T 设为 0.8，阈值 τ 设为 0.3，λ_0 设为 0.4。我们使用 Adam [23]来优化模型中的参数，学习率为$4e$-4，并使用丢弃法，概率为 0.9。此外，我们使用 Recall@K [24]来评估模型性能。Recall@K 是跨模态检索的标准指标，用于计算所有音乐片段中匹配音乐片段的平均命中率，其中音乐片段按照前 K 个匹配分数进行排序。

3.2. 主要结果

表1展示了我们在 MVD 数据集上的实验结果，以及参考 VMR 系统的性能。我们有以下观察：1）与基线 *CMMR* [6]和*CBVMR* [7], 相比，我们提出的框架取得了大幅提升。这是因为*SSVMR* 考虑了对训练极为有害的标签噪声。值得注意的是，*SSVMR* 与*CB-VMR*的参数规模相当，这进一步验证了*SSVMR* 的贡献来自使用 R-Drop 进行半监督自训练，而非参数规模因素。2）尽管*CMVAE* [8] 将相关音乐与视频之间的匹配建模为一个多模态跨模态生成问题，并采用专家乘积原理作为融合策略，*SSVMR*的表现仍然优于它。我们认为，这是因为*SSVMR* 能够捕获关键视频信息，而*CMVAE* 旨在基于这样一个数据集为视频检索音乐：一段音乐片段可能与多个视频相关联。然而，MVD 是一个将一段音乐片段仅与一个视频相关联的数据集，无法为*CMVAE* 充分训练编码器网络提供足够的信息。3）*MRCMV* 能够利用广泛添加的旁白（即未见过的叙述者的声音）有效提升性能。但在未添加旁白时，*SSVMR* 的表现优于它，这验证了我们的框架能够更好地利用原始数据。

模型	R@1	R@10	R@25
CMMR [6]	1.65	6.45	12.70
CBVMR [7, 12]	1.85	7.00	12.90
CMVAE [8]	2.15	10.95	16.30
MRCMV [25]	3.30	12.65	18.20
SSVMR	**4.45**	**15.20**	**24.50**
改进	**34.8%**	**20.2%**	**34.6%**

表1： MVD 数据集上的结果。最佳结果以**粗体**显示。

3.3. 分析

3.3.1. 消融测试

为研究我们方法中各个组件的有效性，我们在 MVD 数据集上进行了消融实验，结果如表2所示。我们移除了反向检索模块，在表2中记为*w/o BR* 。从结果中可以看出，缺少反向检索会导致R@1下降了0.55%，这表明反向检索扩大了数据集，并增强了我们模型的鲁棒性。

模型	R@1
SSVMR	**4.45**
不使用反向检索	3.90(↓ 0.55)
无 *BR* +混合	3.15（↓ 1.30）
无 *BR* +混合 +*R-Drop*	2.50（↓ 1.95）
无 *BR* +混合 +*R-Drop* +*SL*	1.85（↓ 2.60）

表2： MVD上的消融实验结果。

我们进一步移除了依赖显著性模块的混合操作，在表2中记为*w/o BR* + 混合 。我们可以明显观察到 R@1 下降了 1.30%。这表明，在片段级混合两个视频，能够保留更多与基于显著性预测相关的信息，从而提升模型捕捉关键视频信息的能力，使模型能够更好地检索音乐。

我们最终移除R-Drop模块，随后移除自训练模块，分别记为*w/o BR* +混合 + *R-Drop*和*w/o BR* + 混合 + *R-Drop* +*SL* +。R@1同样有所下降，这验证了这些模块有助于处理标签噪声，并且对模型有益。

3.3.2. 替换片段数量的影响

当我们在片段级混合两个视频时，只会替换视频中的一个片段。为了检验只选择一个片段是否足够，我们进行了如第3节所示的一系列实验。我们观察到，随着片段数量增加，R@1逐渐下降。其原因是，替换过多片段会破坏语义，反而不利于模型。

N	R@1
1	**4.45**
2	4.27(↓ 0.18)
3	4.03(↓ 0.42)
4	3.87(↓ 0.58)

表 3： 不同片段数量下 MVD 上的结果。

4. 结论

本文提出了基于**S**显著性**S**自训练的**V**视频-**M**音乐**R**检索（**SSVMR**）方法，将半监督自训练与 R-Drop 相结合。此外，我们依据显著性在片段级混合两个视频，并利用反向检索扩充数据集。实验与分析表明，我们提出的方法能够有效处理标签噪声，增强捕捉关键视频片段的能力，并提升模型的鲁棒性。未来，我们将探索如何在该方法中加入预训练技术，以进一步提升模型性能。

5. 参考文献

[1] Rong Ye、Mingxuan Wang 和 Lei Li，《用于语音翻译的跨模态对比学习》，载于*NAACL* 论文集，2022 年。

[2] Hongxiang Li、Meng Cao、Xuxin Cheng、Zhihong Zhu、Yaowei Li 和 Yuexian Zou，“生成用于视频定位的模板化字幕”，*ArXiv* 预印本，2023 年。

[3] Jinchuan Tian、Brian Yan、Jianwei Yu、Chao Weng、Dong Yu 和 Shinji Watanabe，“贝叶斯风险 CTC：序列到序列任务中可控的 CTC 对齐”，*ArXiv* 预印本，2022 年。

[4] Eric Brochu、Nando De Freitas 和 Kejie Bao，“专辑封面的声音：一种面向多媒体的概率方法”，载于*AISTATS* 会议论文集，2003 年。

[5] Ki-Ho Shin 和 In-Kwon Lee，“利用情感相似度实现视频音乐同步”，大数据与智能计算国际会议，2017 年。

[6]Bochen Li 和 Aparna Kumar，“按视频查询：跨模态音乐检索”，音乐信息检索国际研讨会/会议，2019 年。

[7] Sungeun Hong、Woobin Im 和 Hyun S Yang，“使用软性模态内结构约束的基于内容的视频音乐检索”，载于 *2018* 年 *ACM* 国际多媒体检索会议论文集，2018 年。

[8]Jing Yi、Yaochen Zhu、Jiayi Xie 和 Zhenzhong Chen，“用于基于内容的微视频背景音乐推荐的跨模态变分自编码器”，*IEEE Transactions on Multimedia*，2021 年。

[9] Benoˆıt Fr´enay 和 Michel Verleysen，“存在标签噪声时的分类：综述”，*IEEE Transactions on Neural Networks and Learning Systems*，2013 年。

[10] Sheng Liu、Jonathan Niles-Weed、Narges Razavian 和 Carlos Fernandez-Granda，“早期学习正则化可防止对噪声标签的记忆”，载于*NeurIPS* 论文集，2020 年。[11]Zhihong Zhu、Weiyuan Xu、Xuxin Cheng、Tengtao Song 和 Yuexian Zou，“通过标签语义注入实现的用于口语语言理解的动态图交互框架”，*ArXiv* 预印本，2022 年。

[12]Laure Pr´etet、Ga¨el Richard 和 Geoffroy Peeters，《跨模态音乐视频推荐：设计选择研究》，载于*IJCNN* 论文集，2021 年。

[13] Alexander Schindler 和 Andreas Rauber，《利用音乐相关视觉刻板印象进行音乐信息检索》，*ACMTransactions on IntelligentSystems andTechnology*，2016 年。

[14] Lijun Wu、Juntao Li、Yue Wang、Qi Meng、Tao Qin、Wei Chen、Min Zhang、Tie-Yan Liu 等，《R-drop：用于神经网络的正则化丢弃法》，*NeurIPS* 会议论文集，2021 年。

[15] Rico Sennrich、Barry Haddow 和 Alexandra Birch，《使用单语数据改进神经机器翻译模型》，载于*ACL* 论文集，2016 年。

[16] Chiyuan Zhang、Samy Bengio、Moritz Hardt、Benjamin Recht 和Oriol Vinyals，《理解深度学习（仍然）需要重新思考泛化》，*ACM* 通讯，2021 年。

[17]Eric Arazo、Diego Ortego、Paul Albert、Noel E. O’Connor 和 Kevin McGuinness，《无监督标签噪声建模与损失校正》，载于 *ICML* 论文集，2019 年。

[18] Keyu Tian、Yi Jiang、Qishuai Diao、Chen Lin、Liwei Wang 和 Zehuan Yuan，《为卷积网络设计 BERT：稀疏和分层掩码建模》，*ArXiv* 预印本，2023 年。

[19] Jang-Hyun Kim、Wonho Choo 和 Hyun Oh Song，《Puzzle mix：利用显著性和局部统计量实现最优混合增强》，载于*ICML*论文集，2020 年。

[20] Soyoung Yoon、Gyuwan Kim 和 Kyumin Park，《SS- Mix：用于文本分类的基于显著性的片段混合增强》，载于*ACL Findings*论文集，2021 年。

[21] Xuxin Cheng、Qianqian Dong、Fengpeng Yue、Tom Ko、Mingxuan Wang 和 Yuexian Zou，《M3st：用于语音翻译的三级混合》，*ArXiv*预印本，2022 年。

[22] Sami Abu-El-Haija、Nisarg Kothari、Joonseok Lee、Paul Natsev、George Toderici、Balakrishnan Varadarajan 和 Sudheendra Vijayanarasimhan，《YouTube-8M：大规模视频分类基准》，*ArXiv*预印本，2016 年。

[23] Diederik P. Kingma 和 Jimmy Ba，《Adam：一种随机优化方法》，载于*ICLR* 论文集，2015 年。

[24] Jiwei Wei、Xing Xu、Yang Yang、Yanli Ji、Zheng Wang 和 Heng Tao Shen，《用于跨模态匹配的通用加权度量学习》，载于*CVPR* 论文集，2020 年。

[25] Tingtian Li、Zixun Sun、Haoruo Zhang、Jin Li、Ziming Wu、Hui Zhan、Yipeng Yu 和 Hengcan Shi，《通过利用跨模态编码的旁白实现细粒度视频的深度音乐检索》，载于 *SIGIR* 论文集，2021 年。