合成监督增强的自动音频字幕框架*

肖飞扬[1] 朱乔茜[2] 关 键[1†] 刘徐博[3] 刘濠赫[3]
张可佳[1] 贺广均[4] 王文武[3]

(1 哈尔滨工程大学 计算机科学与技术学院 哈尔滨 150001)
(2 悉尼科技大学 澳大利亚悉尼 NSW 2007)
(3 萨里大学 视觉语音与信号处理中心 英国吉尔福德 GU2 7XH)
(4 天地一体化信息技术国家重点实验室 北京 100095)

2024 年 8 月 7 日收到
2024 年 9 月 30 日定稿

摘要 基于数据驱动的自动音频字幕方法受限于音频–文本数据对的数量和质量，导致其跨模态表示能力不足，制约了整体性能。为此, 提出了一种合成监督增强的自动音频字幕框架 (SynthAC), 该框架利用广泛可用的高质量图像字幕文本语料及文本到音频生成模型生成音频信号, 有效扩充音频–文本数据对, 并通过学习合成音频–文本数据对中的对应关系, 增强音频文本跨模态表示能力。实验表明, 所提 SynthAC 框架通过利用图像字幕中的高质量文本语料库, 显著提升了音频字幕模型性能, 该框架为应对音频–文本数据稀缺挑战提供了有效的解决方案。此外, 该框架可适用于各种主流方法, 在不改变音频字幕模型结构的情况下显著提升音频字幕性能。

关键词 多模态学习, 文本–音频表示, 自动音频字幕, 文本到音频生成
PACS: 43.60, 43.72 **DOI:** 10.12395/0371-0025.2024232 **CSTR:** 32049.14.11-2065.2024232

Enhancing automated audio captioning with synthetic supervision

XIAO Feiyang[1] ZHU Qiaoxi[2] GUAN Jian[1†] LIU Xubo[3] LIU Haohe[3]
ZHANG Kejia[1] HE Guangjun[4] WANG Wenwu[3]

(1 *College of Computer Science and Technology*, *Harbin Engineering University* Harbin 150001)
(2 *University of Technology Sydney* Sydney, Australia NSW 2007)
(3 *Centre for Vision Speech and Signal Processing*, *University of Surrey* Guildford, UK GU2 7XH)
(4 *State Key Laboratory of Space-Ground Integrated Information Technology*, *CAST* Beijing 100095)

Received Aug. 7, 2024
Revised Sept. 30, 2024

Abstract The data-driven automated audio captioning methods are limited by the quantity and quality of available audio-text pairs, resulting in insufficient cross-modal representation, which undermines the captioning performance. To address this, this paper proposes an audio captioning framework enhanced with synthetic supervision, termed SynthAC. This framework leverages commonly available high-quality image captioning text corpus and a text-to-audio generative model to create synthetic audio signals. Therefore, the proposed SynthAC framework can effectively expand audio-text pairs and enhance the cross-modal text-audio representation by learning relations within synthetic audio-text pairs. Experiments demonstrate that the proposed SynthAC framework can significantly improve audio captioning performance by incorporating high-quality text corpus from image captioning, providing an effective solution to the challenge of data scarcity. Additionally, SynthAC can be easily adapted to various state-of-the-art methods, significantly enhancing audio captioning performance without modifying the existing model structures.

Keywords Multimodal learning, Text-audio representation, Automated audio captioning, Text-to-audio generation

* 国家工业和信息化部项目 (CBZ3N21-2) 资助
† 通讯作者: 关键, j.guan@hrbeu.edu.cn

引言

自动音频字幕 (automated audio captioning) 是一种将音频内容转换为文本表述的跨模态任务。不同于仅为音频分配标签的声学场景分类任务[1], 其目的是生成自然语言文本 (即字幕) 以描述音频中包含的声学场景及事件内容[2]。当前, 数据驱动的深度学习方法被广泛用于学习音频和文本信息间的语义关联, 以生成音频字幕[3]。然而, 受采集和标注成本限制, 高质量音频–文本数据稀缺, 现有方法音频文本跨模态表示能力不足, 自动音频字幕性能受限[4]。

当前的解决方案主要包括引入外部数据集[5], 以及利用大语言模型合成字幕文本[6-7]。例如, 网易团队使用大规模人工标注私有数据集训练模型, 在 2021 年声学场景和事件检测和分类挑战任务 6 自动音频字幕赛道获得第一名[5]。然而, 这类私有数据集成本高昂, 且受限于商业许可无法公开, 限制了其在研究和实际应用中的使用。近期, 大语言模型 (如 ChatGPT[8]) 被用于扩充音频字幕数据集的字幕文本数量[6-7]。例如, 使用 ChatGPT 将弱标注的音频标签转换为流畅的字幕文本, 获得了音频字幕数据集 WavCaps, 增加了音频–文本数据对的数量, 提升了自动音频字幕性能[6]。此外, 基于 ChatGPT 的混合增强策略通过 ChatGPT 混合音频字幕数据集中的任意两条字幕文本内容, 生成全新的字幕文本, 扩充音频–文本数据对并提高模型性能[7]。

上述研究通过合成字幕文本扩充音频字幕数据集, 但仍受限于可用的音频数据量。本文考虑到具有场景描述信息的文本语料库丰富且易于获取, 而且文本语料通常兼顾了视觉与声学的语义内容。例如, “一只吠叫的狗”不仅传达了视觉信息“吠叫的狗”, 还包含了声学场景语义内容“狗吠”。因此, 其他跨模态领域的文本描述信息可能可用于音频文本跨模态任务。

故本文以文本描述为切入点, 合成音频信号而非字幕文本, 从而扩充音频–文本数据对, 缓解数据稀缺问题。通过文本语料库和文本到音频生成技术, 合成音频信号, 扩充音频–文本数据对用于训练模型, 提升自动音频字幕模型性能。文本到音频生成技术能够从文本中提取声学场景与事件相关的语义内容, 生成高质量音频数据[9]。合成的音频–文本数据与现有数据集结合, 采用监督学习训练模型, 从而提升其性能。这种将真实与合成数据结合, 共同用于监督学习训练的策略即为合成监督策略。近年来, 该策略已成功应用于异音检测[10]、唇读识别[11]、语音识别[12] 等领域, 有效缓解了数据稀缺问题并提升了模型精度。

在音频–文本多模态领域中, 合成监督策略之所以能够有效解决数据稀缺问题并增益模型性能, 主要原因在于文本到音频生成技术能够挖掘文本中包含的场景语义, 生成对应的合成音频, 从而扩充声学场景语义多样性。大量研究表明, 文本到音频生成模型能够通过跨模态对齐和转换, 将文本描述中的声学场景语义映射为音频内容。例如, 文本到一般音频生成 (Text-to-Audio, TTA) 领域的 AudioLDM[9]、AudioLDM 2[13] 等方法、文本生成语音 (Text-to-Speech, TTS) 领域的 DiffVoice[14]、ProDiff[15] 等方法、以及文本生成音乐 (Text-to-Music, TTM) 领域的 Music ControlNet[16]、MusicLDM[17] 等方法, 均表明了文本到音频生成模型能够令文本语义以听觉形式被感知和处理。

基于此, 本文提出了一种合成监督增强的自动音频字幕框架 (automated audio captioning with synthetic supervision, SynthAC), 通过文本语料库和文本到音频生成技术, 扩充音频–文本数据对, 构建合成监督训练集以提升模型性能。本文使用图像字幕数据集 COCO 中的图像字幕文本数据[18-19], 通过文本到音频生成模型 AudioLDM[9] 挖掘图像字幕文本中潜在的声学场景和事件信息, 生成具有相应内容的合成音频, 增加了合成的音频–文本数据对的声学场景语义多样性。然后, 将合成的音频–文本数据对与现有数据集 (如 AudioCaps[20]) 合并, 以构建合成监督训练集用于训练模型, 使得模型学习到更丰富的音频–文本特征信息。该框架增强了模型的跨模态表示能力, 提升了自动音频字幕性能, 为该领域带来了新的解决方案。

SynthAC 框架不仅提升了自动音频字幕模型性能, 还减少了对真实音频–文本数据的依赖, 显著增强了跨模态表示能力。与 SpecAugment[21] 等数据增广方法相比, SynthAC 框架通过引入外部文本语料中的声学场景语义信息, 增加了数据多样性。而 SpecAugment 仅在已有数据上进行特征变换, 无法增加语义信息, 因此在提升模型性能方面存在局限。相比之下, SynthAC 框架通过引入全新的声学场景语义, 构建了更具多样性的训练集, 提升了自动音频字幕性能。

为验证 SynthAC 框架的有效性, 本文将其用于两种主流自动音频字幕模型 (GraphAC[22] 和 P-Transformer[23]), 并与主流方法在 AudioCaps 数据集[20]

上进行对比实验。结果表明, SynthAC 框架在不改变模型结构的情况下, 能够显著提升自动音频字幕性能。此外, 在 AudioCaps 数据量减半甚至不足一半的情况下, SynthAC 框架仍表现出优于现有主流方法的性能, 证明了其在解决自动音频字幕的数据稀缺问题上的潜力。本文所提框架生成的合成音频信号示例可在如下链接获取: https://github.com/LittleFlyingSheep/SynthAC。

1 基于合成监督的自动音频字幕框架整体流程

所提基于合成监督的自动音频字幕框架 SynthAC 的整体流程如图 1 所示。本文采用高质量标注的图像字幕 (来自 COCO 数据集的字幕文本[18]) 作为文本到音频生成模型 AudioLDM[9] 的输入条件, 用以获取合成音频信号及构建合成数据集。随后, 将合成数据集与现有音频字幕数据集 (AudioCaps[20]) 合并, 构建合成监督训练集, 用于自动音频字幕模型 (GraphAC[22]) 训练, 以增强其跨模态表示学习能力并提高自动音频字幕性能。需要注意的是, 除 GraphAC 之外, 所提 SynthAC 框架还可以适用于其他自动音频字幕模型。

1.1 基于图像字幕文本驱动的音频合成

为了获得合成音频信号以扩充音频–文本数据对的数量, 本文将图像字幕文本用于文本到音频生成模型 AudioLDM, 以挖掘图像字幕中潜在的声学场景语义内容, 生成相对应的合成音频信号。为说明图像字幕用于文本到音频生成的合理性, 本文提供了如表 1 的示例分析, 展示了图像字幕与音频字幕之间共通的语义内容, 以阐明图像字幕与声学场景及事件语义内容存在潜在关联。

如表 1 所示, 高质量的图像字幕文本所描述的视觉场景内容中, 通常包含能够发出声音的目标实体。这些声音对应某种声学场景或事件的语义内容。例如, 在示例 1 中, 图像字幕文本提到了“吠叫的犬 (a barking dog)”, 该目标实体对应的声音便是“犬吠 (dog barking)”这一声学场景的语义内容。因此, 基于这一潜在关联, 在图像字幕文本驱动的音频合成阶段, 本文能够利用文本到音频生成模型挖掘图像字幕中的潜在声学场景语义内容, 从而获取对应的合成音频信号。

具体而言, 所提框架首先采用基于对比学习的音频文本预训练模型 (contrastive language-audio pretraining, CLAP)[24], 从高质量标注的图像字幕文本 t_{img} 中提取文本嵌入向量, 该过程可表示为

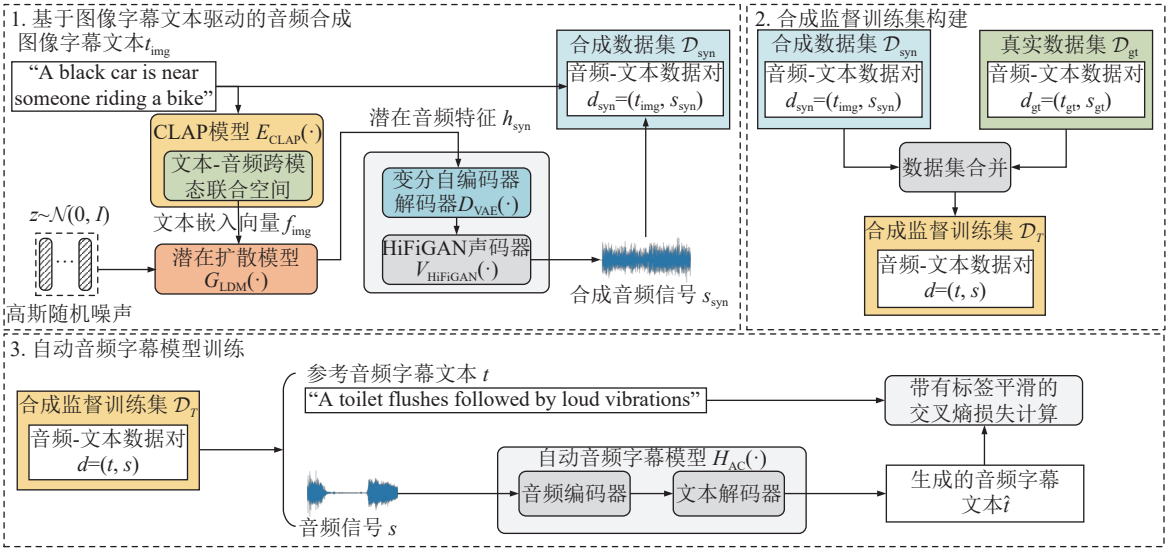

图 1　SynthAC 框架的总体示意图

表 1　图像字幕文本与音频字幕文本之间语义内容潜在关联的示例说明

示例序号	图像描述文本	音频描述文本	共有的语义内容
1	**A barking dog** looks over a ledge lined with Christmas lights	**Dog barking** and growling	Dog barking
2	A black **car** is near someone riding a bike	A man talking and a **car** passing by loudly	Car
3	**A cat sleeping** on a rock near a bike	**A cat sleeps** and snores	A cat sleeps

$$f_{\text{img}} = E_{\text{CLAP}}(t_{\text{img}}), \tag{1}$$

其中, $E_{\text{CLAP}}(\cdot)$表示 CLAP 模型, f_{img}表示图像字幕文本 t_{img}对应的文本嵌入向量。由于 CLAP 模型能够从其文本–音频跨模态联合空间中提取声学场景语义内容相关的文本嵌入向量, 上述过程能够有效捕获图像字幕文本中所蕴含的声学场景语义内容, 为高质量音频合成提供支撑。

然后, 在图像字幕文本驱动的音频合成阶段, 借助文本嵌入向量蕴含的声学场景语义内容, 指导文本到音频生成模型 AudioLDM[9] 生成音频信号。在文本嵌入向量 f_{img}的指导下, AudioLDM 中的潜在扩散模型 (latent diffusion model, LDM) 能够将采样自高斯分布的随机噪声 z转换为与文本嵌入向量中声学场景内容信息相对应的潜在音频特征 h_{syn}。该过程可表示为

$$h_{\text{syn}} = G_{\text{LDM}}(f_{\text{img}}, z), \tag{2}$$

其中, $G_{\text{LDM}}(\cdot)$表示 AudioLDM 中的潜在扩散模型。这一转换过程至关重要, 因为其建立了文本语义内容和音频听觉表达之间的信息关联, 实现了从文本模态到音频模态的数据特征分布转化, 使得通过合成音频信号解决自动音频字幕的数据稀缺问题成为可能。

随后, 潜在音频特征 h_{syn}会通过 AudioLDM 模型中的变分自编码解码器 (variational auto-encoder decoder, VAE decoder)[9] 和 HiFiGAN 声码器 (HiFiGAN vocoder)[25], 转换为音频波形信号作为输出的合成音频信号结果。这一过程可表示为

$$s_{\text{syn}} = V_{\text{HiFiGAN}}(D_{\text{VAE}}(h_{\text{syn}})), \tag{3}$$

其中, $D_{\text{VAE}}(\cdot)$表示 AudioLDM 模型中的变分自编码解码器, $V_{\text{HiFiGAN}}(\cdot)$则表示 AudioLDM 模型中的 HiFiGAN 声码器, s_{syn}表示最终获得的合成音频信号。在这一过程中, 变分自编码解码器用于将潜在音频特征转换为对数梅尔谱图特征, 而 HiFiGAN 声码器则用于将对数梅尔谱图特征转换为高质量的合成音频信号。

通过这种方式, 即可获得基于图像字幕文本与合成音频信号组合而成的音频–文本数据对 $d_{\text{syn}} = (t_{\text{img}}, s_{\text{syn}})$。而所有来自于图像字幕文本与合成音频信号的音频–文本数据对, 则共同构建成一个合成数据集 $\mathcal{D}_{\text{syn}}$, 用于后续的合成监督训练集构建阶段。

1.2 合成监督训练集构建

在合成监督训练集构建阶段, 将图像字幕文本驱动的音频合成阶段获得的合成数据集 $\mathcal{D}_{\text{syn}}$, 与现有的真实数据集 $\mathcal{D}_{\text{gt}}$ (例如 AudioCaps) 合并, 以扩充音频–文本数据对的数量, 构建成合成监督训练集 $\mathcal{D}_T$。这一操作可表示为

$$\mathcal{D}_T = \mathcal{D}_{\text{syn}} \cup \mathcal{D}_{\text{gt}}, \tag{4}$$

其中, $\mathcal{D}_T$表示合成监督训练集, 合成监督训练集中的音频–文本数据对可表示为 $d = (t, s)$, t表示字幕文本, s表示与字幕文本对应的音频信号。此时, 所有音频信号均有对应的字幕文本作为参考标签, 因此该训练集是有监督条件下的数据集, 又由于其构建过程中引入了合成数据集, 故称其为合成监督训练集。

1.3 自动音频字幕模型训练

在自动音频字幕模型训练阶段, 所提框架选择了自动音频字幕模型 GraphAC 模型[22] 进行训练, 构成的自动音频字幕方法称为 Synth-GraphAC。GraphAC 模型采用自编码器结构, 其中音频编码器用于提取包含声学场景及事件信息的音频特征, 文本解码器用于从音频特征中解码语义内容信息, 标注音频字幕文本。GraphAC 模型的音频编码器引入了图注意力模块[26], 以进一步捕捉音频特征提取模块 (即 PANNs 模块[27]) 输出特征中的上下文时序信息。文本解码器使用了一个带有 Word2Vec 语言模型[28] 的两层 Transformer 模块[29] 来生成音频字幕文本。

在模型训练过程中, 音频信号 s被输入到自动音频字幕模型中, 以标注音频字幕文本。该过程可表示为

$$\hat{t} = H_{\text{AC}}(s), \tag{5}$$

其中, $H_{\text{AC}}(\cdot)$表示自动音频字幕模型, 在 Synth-GraphAC 方法中为 GraphAC 模型, $\hat{t}$表示生成的音频字幕文本。

随后, 通过带有标签平滑的交叉熵损失函数[22,23] 优化所提框架中音频字幕模型, 损失函数的公式为

$$\mathcal{L} = -\sum_i \left[(1-\epsilon)\cdot t_i + \frac{\epsilon}{k}\right]\log(\hat{t}_i), \tag{6}$$

其中, $\mathcal{L}$表示带有标签平滑的交叉熵损失函数值, k表示语言模型词库中的词汇总数, ϵ则表示平滑系数, 用于平衡不同词汇的词频差异, 缓解词频不均衡现象对自动音频字幕结果的影响。

综上所述, 本文提出的 SynthAC 框架通过合成音频信号, 构建合成监督训练集训练音频字幕模型, 从而增强模型的文本–音频跨模态表示能力, 提升自动音频字幕性能。值得注意的是, 除 GraphAC 模型

外, 所提 SynthAC 框架还适用于其他音频字幕模型, 以提高其性能。在实验验证中, 本文所提的 SynthAC 框架还被应用于另一种自动音频字幕模型, 即 P-Transformer[23], 并将该方法命名为 Synth-P-Transformer。实验证明了所提框架在 GraphAC 与 P-Transformer 上均能提升自动音频字幕性能, 说明了所提框架的有效性与通用性, 详见 3.1 节。

2 实验设计

2.1 数据集

为实现所提框架中图像字幕文本驱动的音频合成, 本文采用了视觉文本跨模态领域广泛使用的图像字幕数据集 COCO[18], 从中抽取图像字幕文本用于音频合成。COCO 数据集提供了 414113 条人工标注的高质量图像字幕文本, 用于描述视觉场景内容。本文从 COCO 数据集中随机选择了 25000 条图像字幕文本, 作为 AudioLDM 模型的输入文本, 生成了对应的 25000 条合成音频信号。这些合成音频信号用于构建合成监督训练集, 从而增强自动音频字幕模型的文本–音频跨模态表示能力。

为实现自动音频字幕模型训练, 本文采用了自动音频字幕领域广泛使用的音频字幕数据集 AudioCaps[20], 该数据集提供了 51744 组音频文本数据对。遵循主流方法的数据集划分设置, 本文将 AudioCaps 数据集的开发集与验证集合并作为训练阶段所用的真实数据集, 利用其评估集进行自动音频字幕模型的性能评估[22,23,30]。在实验中, 音频信号的采样率统一设置为 16 kHz。

2.2 评价指标

为评估自动音频字幕模型的性能表现, 本文采用了主流方法所使用的评价指标体系, 包含词级别的评价指标 BLEU_n[31]、ROUGE_l[32] 和 METEOR[33], 与语义级别的评价指标 CIDEr[34]、SPICE[35]、SPIDEr[36] 和 SPIDEr-FL[37]。

其中, 词级别的评价指标 BLEU_n、ROUGE_l 和 METEOR 用于衡量生成的音频字幕文本与参考音频字幕文本在词汇层面的匹配度[23]。需要注意的是, BLEU_n 指标评价的是 n 个连续词之间的匹配度, 本文评价了 BLEU_1、BLEU_2、BLEU_3 和 BLEU_4 这 4 种情况的 BLEU_n 指标表现。ROUGE_l 指标则侧重于生成的音频字幕文本与参考音频字幕文本的最长公共子序列匹配程度。METEOR 指标则考虑了词形变化与同义词匹配情况, 相比前两者更具鲁棒性。

语义级别的评价指标 CIDEr 衡量生成的音频字幕文本的流畅性[34]。SPICE 衡量模型生成的音频字幕文本中声学场景语义内容的完整程度[35]。SPIDEr 是 CIDEr 和 SPICE 指标的平均值, 兼顾了评价过程对文本流畅性和语义内容准确性的评价[36]。SPIDEr-FL 在 SPIDEr 的基础上引入了基于流畅性误差的惩罚, 进一步关注生成的音频字幕文本的句式表达, 提高了评估的鲁棒性[37]。因此, SPIDEr-FL 指标是语义级别指标中最为重要的评价指标。

2.3 实验参数

在基于图像字幕文本驱动的音频合成阶段, 所提框架使用了 AudioLDM 模型的“audioldm-l-full”版本[9] 进行音频合成。合成音频信号的时长统一设置为 10 s, 与 AudioCaps 数据集中的音频信号时长保持一致。在自动音频字幕模型训练阶段, Synth-GraphAC 和 Synth-P-Transformer 的批处理大小 (batch size) 均设置为 16, 采用 AdamW 优化器[38] 更新模型参数, 学习率设置为 0.001。此外, 为了增强自动音频字幕模型的泛化能力, 本文遵循主流方法的设计, 在音频字幕模型训练中使用了谱图特征增强策略 SpecAugment[21], 以增加音频特征的多样性, 增强自动音频字幕模型的音频特征表示能力[22,23,30]。

3 对比实验与分析

本文的对比实验如下: 首先, 通过与主流自动音频字幕模型的性能对比, 验证了所提 SynthAC 框架的有效性, 并说明了所提框架可适用于不同的自动音频字幕模型, 具备一定通用性。随后, 进一步比较了所提框架在不同数据量下的性能, 以讨论其在数据稀缺条件下的适用性和优势。通过这些分析, 本文旨在全方面展示所提框架对自动音频字幕研究的增益效果和应用潜力。

3.1 有效性

为了验证所提 SynthAC 框架的有效性, 表 2 和表 3 给出了基于所提框架的方法与仅使用真实数据训练的主流方法的对比实验结果。对比的主流方法包括 GPT-Similar[39]、TopDown-AlignedAtt[20]、P-Transformer[23]、GraphAC[22] 和 P-LocalAFT[30]。其中, GPT-Similar 是一种使用相似文本检索标注音频字幕文本的方法[39]。TopDown-AlignedAtt 是一种基于注意力机制的音频字幕方法[20]。P-Transformer 是一种基于预训练音频特征表示模型的方法[23]。GraphAC 和 P-LocalAFT 则是在 P-Transformer 方法的基础上

表 2 AudioCaps 评估集上的词级别指标性能

自动音频字幕方法	$BLEU_1$ (%)	$BLEU_2$ (%)	$BLEU_3$ (%)	$BLEU_4$ (%)	$ROUGE_l$ (%)	METEOR (%)
GPT-Similar[39]	63.8	45.8	31.8	20.4	43.4	19.9
TopDown-AlignedAtt[20]	61.4	44.6	31.7	21.9	45.0	20.3
P-LocalAFT[30]	66.0	47.9	34.6	24.6	46.4	22.3
P-Transformer[23]	53.4	38.9	27.1	18.0	44.2	21.5
Synth-P-Transformer	**67.7**	**49.9**	**36.0**	**25.1**	**46.8**	**22.7**
GraphAC[22]	64.5	47.8	34.3	23.7	46.1	**22.4**
Synth-GraphAC	**66.5**	**48.7**	**35.2**	**24.7**	**46.4**	**22.4**

表 3 AudioCaps 评估集上的语义级别指标性能

自动音频字幕方法	CIDEr (%)	SPICE (%)	SPIDEr (%)	SPIDEr-FL (%)
GPT-Similar[39]	50.3	13.9	32.1	—
TopDown-AlignedAtt[20]	59.3	14.4	36.9	—
P-LocalAFT[30]	64.1	16.6	40.4	40.0
P-Transformer[23]	57.7	16.6	37.1	35.9
Synth-P-Transformer	**63.9**	**16.7**	**40.3**	**39.4**
GraphAC[22]	64.4	**16.7**	40.5	39.3
Synth-GraphAC	**65.6**	16.5	**41.0**	**40.4**

发展而来, 分别加强了对音频特征时序上下文信息与局部事件信息的关注, 取得了一定性能提升[22,30]。此外, 为了验证所提框架的通用性, 本文在 SynthAC 框架中分别使用了 P-Transformer 和 GraphAC 作为自动音频字幕模型, 得到自动音频字幕方法 Synth-P-Transformer 和 Synth-GraphAC, 用于性能对比验证。

表 2 和表 3 分别给出了本文构建的 Synth-P-Transformer 和 Synth-GraphAC 方法与其他对比方法在词级别指标和语义级别指标上的性能表现。通过表 2 可以发现, 本文所构建的 Synth-P-Transformer 和 Synth-GraphAC 方法在词级别评价指标上均明显优于对比方法。从表 3 可知, 在语义级别的评价指标上, 本文构建的 Synth-P-Transformer 方法 SPIDEr-FL 指标评价得分仅次于 P-LocalAFT, 而 Synth-GraphAC 方法在 CIDEr、SPIDEr 和 SPIDEr-FL 优于所有对比方法。这说明本文构建的 Synth-P-Transformer 和 Synth-GraphAC 方法在自动音频字幕的词级别与语义级别评价指标上, 均取得了较为出色的表现。上述结果表明, 所提框架是一种行之有效的自动音频字幕解决方案。

此外, 根据表 2 和表 3 中 Synth-GraphAC 方法与 GraphAC 方法的评价指标性能比较, 以及 Synth-P-Transformer 与 P-Transformer 的评价指标性能比较, 可以发现 Synth-GraphAC 方法在所有词级别指标与语义级别指标 CIDEr、SPIDEr 和 SPIDEr-FL 上均优于 GraphAC 方法, 而 Synth-P-Transformer 方法则是在所有评价指标表现上均优于 P-Transformer 方法。这说明基于所提框架的方法的整体性能表现优于不使用所提框架的方法。由于二者之间的差别仅在于是否使用所提框架, 而无其他差异, 因此上述结果说明, 本文所提出的 SynthAC 框架可以在不改变自动音频字幕模型结构的情况下, 显著提高自动音频字幕的性能。这表明本文所提框架具备一定的通用性, 能够适用于不同的自动音频字幕模型, 提升其性能表现。

需要注意的是, 由于 SynthAC 框架引入了基于合成音频的音频–文本数据对扩充了训练数据量, 故相比未采用 SynthAC 框架的自动音频字幕模型 (如 P-Transfomer 和 GraphAC) 训练时间略有增加, 所增加的训练时间与增加的音频–文本数据对数量成正比。测试阶段由于测试集固定不变, 故基于 SynthAC 框架的方法测试时间与原本的测试时间相同。

为进一步说明所提框架对自动音频字幕的提升效果, 本文提供了如表 4 所示的 3 组音频字幕文本示例。通过对比 Synth-P-Transformer 与 P-Transformer 方法生成的音频字幕文本, 说明所提框架在自动音频字幕结果上的提升。如表 4 所示, 示例 1 中 P-Transformer 方法将“blades running”(桨叶运转声) 错误地解释为了“gun fires rapidly”(重复枪击声), 而 Synth-P-Transformer 方法则通过所提框架增强的文本–音频表示能力准确地描述了这一概念。在示例 2 中, Synth-P-Transformer 准确描述了音频事件之间的上下文信息, 即“followed by”(紧接着), 并且生成的音频字幕文本与参考音频字幕文本完全一致, 而 P-

表 4　使用与不使用所提框架时生成的音频字幕文本示例

示例序号	示例来源	是否使用所提框架	音频字幕文本
1	评估集参考字幕文本	否	A helicopter **blades running**
	P-Transformer[23]	否	A helicopter machine *gun fires rapidly*
	Synth-P-Transformer	**是**	Helicopter **blades spinning**
2	评估集参考字幕文本	否	A man talking **followed by** a toilet flushing
	P-Transformer[23]	否	A man speaking a toilet flushing
	Synth-P-Transformer	**是**	A man speaks **followed by** a toilet flushing
3	评估集参考字幕文本	否	A woman speaks with some rattling and some **spraying**
	P-Transformer[23]	否	An adult female is speaking
	Synth-P-Transformer	**是**	A woman speaking followed by **spraying**

Transformer 方法则忽略了音频事件之间的上下文关系。示例 3 中 P-Transformer 方法未能捕获到“spraying”(喷洒声) 这一声学事件, 而 Synth-P-Transformer 方法则正确地表述了这一声学事件。这些结果进一步支撑了表 2 和表 3 中基于所提框架的自动音频字幕方法在词级别各项指标、语义级别的流畅度指标 CIDEr 以及核心语义指标 SPIDEr-FL 等方面的提升。

综上所述, 本小节对比实验结果显示, 基于所提框架的自动音频字幕方法表现出优于主流方法的自动音频字幕性能, 说明了本文所提策略的有效性, 即利用文本到音频生成模型, 挖掘文本语料中包含的声学场景及事件内容, 合成音频数据用于音频字幕模型训练, 能够显著提升自动音频字幕模型的性能表现。此外, 实验结果还表明, 本文所提框架具备一定通用性, 能够在不改变自动音频字幕模型结构的情况下, 只通过引入额外的基于合成音频信号的音频–文本数据对进行模型训练, 提升其自动音频字幕性能。因此, 所提框架能够推广到不同的自动音频字幕模型, 提升其性能表现。

3.2　不同数据量下的性能比较

为进一步验证所提框架在数据稀缺场景下的表现, 本节对比实验使用了不同数量的真实音频–文本数据对, 对不使用所提框架的方法 (即 P-Transformer) 和基于所提框架的方法 (即 Synth-P-Transformer) 进行训练, 并对二者的评价指标表现进行对比分析。在实验中, 分别采用了 AudioCaps 数据集 12.5%、25.0%、37.5% 和 50.0% 数据量的音频–文本数据对用于训练, 实验结果如表 5 所示。

如表 5 所示, 本文所提框架能够显著提升自动音频字幕模型在使用不同数量真实音频–文本数据对进行训练时的性能, 尤其是在数据量非常有限的情况下。例如: 仅使用 12.5% 的 AudioCaps 数据集时, Synth-P-Transformer 方法的表现已经优于表 2 和表 3 所示的 GPT-Similar 方法。在使用仅 37.5% 的 AudioCaps 数据集时, 其音频字幕性能甚至优于使用完整 AudioCaps 数据集训练的 P-Transformer 方法。实验结果验证了所提框架利用外部文本生成音频用于合成监督的策略, 在音频–文本数据对数量有限条件下的有效性, 表明所提框架为自动音频字幕研究中的数据稀缺挑战提供了一种可行的解决方案。

综上所述, 本文实验验证了通过合成音频信号扩充可用音频–文本数据对, 增强自动音频字幕模型文本–音频跨模态表示能力的可行性, 为解决数据稀缺挑战提供了一种行之有效的通用解决方案。此外, 本文所提框架的实验结果表明, 合理利用其他跨模态领域 (如视觉文本跨模态领域) 的高质量标注文本信息, 有助于解决音频文本跨模态领域下游任务 (如自动音频字幕) 中的数据稀缺问题。

表 5　不同数据量下的使用与不使用所提框架时的自动音频字幕性能表现

AudioCaps所用数据量	是否使用所提框架	METEOR (%)	CIDEr (%)	SPICE (%)	SPIDEr (%)	SPIDEr-FL (%)
12.5%	否	19.4	49.9	14.3	32.1	29.9
	是	**20.7**	**55.2**	**15.1**	**35.2**	**33.1**
25.0%	否	19.9	56.1	13.6	34.8	33.8
	是	**21.3**	**58.4**	**15.5**	**36.9**	**35.0**
37.5%	否	21.0	58.1	15.1	36.6	35.3
	是	**21.8**	**61.1**	**15.9**	**38.5**	**37.6**
50.0%	否	21.2	57.6	16.2	36.9	34.4
	是	**22.5**	**63.8**	**16.7**	**40.2**	**38.3**

4 结论

本文提出了一种基于合成监督的自动音频字幕框架, 该框架利用图像字幕数据集中高质量标注的字幕文本和文本到音频生成模型, 获取合成音频信号, 构建合成监督训练集用于模型训练, 从而增强音频字幕模型的跨模态表示能力, 提高自动音频字幕的性能。实验结果表明, 基于所提框架构建的自动音频字幕方法在词级别与语义级别指标上均取得了优于主流方法的性能表现。此外, 所提框架具备一定通用性, 可以推广到不同的自动音频模型。即使在音频–文本训练数据极其有限的情况下, 所提框架仍能保持出色的性能表现, 为解决自动音频字幕中的数据稀缺问题提供了可行的解决方案。在后续研究中, 可以进一步探索多种声学场景事件混合的复杂声学场景下的合成音频信号生成, 以帮助模型更好地应对实际场景的多样性和复杂性, 从而进一步提升自动音频字幕模型的性能和泛化能力。

参 考 文 献

1 姚琨, 杨吉斌, 张雄伟, 等. 基于多分辨率时频特征融合的声学场景分类. 声学技术, 2020; **39**(4): 494−500

2 Drossos K, Lipping S, Virtanen T. Clotho: An audio captioning dataset. International Conference on Acoustics, Speech and Signal Processing, IEEE, Barcelona, Spain, 2020: 736−740

3 陈耕耘, 李圣辰, 邵曦, 等. 基于迁移学习与强化学习的自动音频标注系统. 复旦学报 (自然科学版), 2022; **61**(5): 520−526

4 Liu X, Mei X, Huang Q, *et al*. Leveraging pre-trained BERT for audio captioning. European Signal Processing Conference, IEEE, Belgrade, Serbia, 2022: 1145−1149

5 Yuan W, Han Q, Liu D, *et al*. The DCASE 2021 challenge task 6 system: Automated audio captioning with weakly supervised pre-training and word selection methods. DCASE2021 Challenge, 2021

6 Mei X, Meng C, Liu H, *et al*. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2024; **32**: 3339−3354

7 Wu S L, Chang X, Wichern G, *et al*. Improving audio captioning models with fine-grained audio features, text embedding supervision, and LLM mix-up augmentation. International Conference on Acoustics, Speech and Signal Processing, IEEE, Seoul, South Korea, 2024: 316−320

8 OpenAI, Achiam J, Adler S, *et al*. GPT-4 technical report. arXiv preprint: 2303.08774, 2024

9 Liu H, Chen Z, Yuan Y, *et al*. AudioLDM: Text-to-audio generation with latent diffusion models. International Conference on Machine Learning, PMLR, Hawaii, USA, 2023: 21450−21474

10 Zhang H, Zhu Q, Guan J, *et al*. First-shot unsupervised anomalous sound detection with unknown anomalies estimated by metadata-assisted audio generation. International Conference on Acoustics, Speech and Signal Processing, IEEE, Seoul, South Korea, 2024: 1271−1275

11 Liu X, Lakomkin E, Vougioukas K, *et al*. SynthVSR: Scaling up visual speech recognition with synthetic supervision. Conference on Computer Vision and Pattern Recognition, IEEE/CVF, Vancouver, Canada, 2023: 18806−18815

12 Fazel A, Yang W, Liu Y, *et al*. SynthASR: Unlocking synthetic data for speech recognition. Interspeech Conference, ISCA, Brno, Czechia, 2021: 896−900

13 Liu H, Yuan Y, Liu X, *et al*. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2024; **32**: 2871−2883

14 Liu Z, Guo Y, Yu K. DiffVoice: Text-to-speech with latent diffusion. International Conference on Acoustics, Speech and Signal Processing, IEEE, Rhodes Island, Greece, 2023: 1−5

15 Huang R, Zhao Z, Liu H, *et al*. ProDiff: Progressive fast diffusion model for high-quality text-to-speech. ACM International Conference on Multimedia, ACM, Lisbon, Portugal, 2022: 2595−2605

16 Wu S L, Donahue C, Watanabe S, *et al*. Music ControlNet: Multiple time-varying controls for music generation. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2024; **32**: 2692−2703

17 Chen K, Wu Y, Liu H, *et al*. MusicLDM: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. International Conference on Acoustics, Speech and Signal Processing, IEEE, Seoul, South Korea, 2024: 1206−1210

18 Lin T Y, Maire M, Belongie S, *et al*. Microsoft COCO: Common objects in context. European Conference on Computer Vision, Springer, Zurich, Switzerland, 2014: 740−755

19 佟国香, 李乐阳. 基于图神经网络和引导向量的图像字幕生成模型. 数据采集与处理, 2023; **38**(1): 209−219

20 Kim C D, Kim B, Lee H, *et al*. AudioCaps: Generating captions for audios in the wild. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACL, Minneapolis, USA, 2019: 119−132

21 Park D S, Chan W, Zhang Y, *et al*. SpecAugment: A simple data augmentation method for automatic speech recognition. Interspeech Conference, ISCA, Graz, Austria, 2019: 2613−2617

22 Xiao F, Guan J, Zhu Q, *et al*. Graph attention for automated audio captioning. *IEEE Signal Process. Lett.*, 2023; **30**: 413−417

23 Mei X, Huang Q, Liu X, *et al*. An encoder-decoder based audio captioning system with transfer and reinforcement learning. Detection and Classification of Acoustic Scenes and Events Workshop, IEEE, Online, 2021: 206−210

24 Wu Y, Chen K, Zhang T, *et al*. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. International Conference on Acoustics, Speech and Signal Processing, IEEE, Rhodes Island, Greece, 2023: 1−5

25 Kong J, Kim J, Bae J. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems, Curran Associates, Online, 2020: 17022−17033

26 郝超, 裘杭萍, 孙毅. 融合 BERT 和图注意力网络的多标签文本分类. 计算机系统应用, 2022; **31**(6): 167−174

27 Kong Q, Cao Y, Iqbal T, *et al*. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2020; **28**: 2880−2894

28 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Advances

in Neural Information Processing Systems, Curran Associates, Harrahs and Harveys, USA, 2013: 3111−3119

29 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Advances in Neural Information Processing Systems, Curran Associates, Long Beach Convention Center, USA, 2017: 6000−6010

30 Xiao F, Guan J, Lan H, *et al*. Local information assisted attention-free decoder for audio captioning. *IEEE Signal Process. Lett.*, 2022; **29**: 1604−1608

31 Papineni K, Roukos S, Ward T, *et al*. BLEU: A method for automatic evaluation of machine translation. Annual Meeting on Association for Computational Linguistics, ACL, Philadelphia, USA, 2002: 311−318

32 Lin C Y. ROUGE: A package for automatic evaluation of summaries. Text Summarization Branches Out, ACL, Barcelona, Spain, 2004: 74−81

33 Banerjee S, Lavie A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, ACL, Ann Arbor, USA, 2005: 65−72

34 Vedantam R, Zitnick C L, Parikh D. CIDEr: Consensus-based image description evaluation. Conference on Computer Vision and Pattern Recognition, IEEE/CVF, Boston, MA, USA, 2015: 4566−4575

35 Anderson P, Fernando B, Johnson M, *et al*. SPICE: Semantic propositional image caption evaluation. European Conference on Computer Vision, Springer, Amsterdam, The Netherlands, 2016: 382−398

36 Liu S, Zhu Z, Ye N, *et al*. Improved image captioning via policy gradient optimization of SPIDEr. International Conference on Computer Vision, IEEE, Venice, Italy, 2017: 873−881

37 Zhou Z, Zhang Z, Xu X, *et al*. Can audio captions be evaluated with image caption metrics? International Conference on Acoustics, Speech and Signal Processing, IEEE, Singapore, 2022: 981−985

38 Loshchilov I, Hutter F. Decoupled weight decay regularization. International Conference on Learning Representations, OpenReview, New Orleans, Louisiana, USA, 2018

39 Koizumi Y, Ohishi Y, Niizumi D, *et al*. Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval. arXiv preprint: 2012.07331, 2020