计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
62(8): 1875−1883, 2025

大语言模型知识引导的开放域多标签动作识别

朱荣江[1,2] 石语珩[1,2] 杨 硕[3,4] 王子奕[1,2] 吴心筱[1,2]

[1]（北京理工大学计算机学院 北京 100081）

[2]（智能信息技术北京市重点实验室（北京理工大学） 北京 100081）

[3]（深圳北理莫斯科大学 广东深圳 518172）

[4]（广东省智能感知与计算重点实验室（深圳北理莫斯科大学） 广东深圳 518172）

（3220241447@bit.edu.cn）

Open-Vocabulary Multi-Label Action Recognition Guided by LLM Knowledge

Zhu Rongjiang[1,2], Shi Yuheng[1,2], Yang Shuo[3,4], Wang Ziyi[1,2], and Wu Xinxiao[1,2]

[1]（*School of Computer Science & Technology*, *Beijing Institute of Technology*, *Beijing* 100081）

[2]（*Beijing Key Laboratory of Intelligent Information Technology* (*Beijing Institute of Technology*), *Beijing* 100081）

[3]（*Shenzhen MSU-BIT University*, *Shenzhen*, *Guangdong* 518172）

[4]（*Guangdong Key Laboratory of Machine Perception and Intelligent Computing* (*Shenzhen MSU-BIT University*), *Shenzhen*, *Guangdong* 518172）

Abstract Open-vocabulary multi-label action recognition tasks aim to identify various human actions in videos that are not seen during the training phase. Compared with traditional action recognition, this task is more practical as it closely mirrors real-world scenarios and has broader application prospects. However, it poses significant challenges in effectively generalizing models to unseen action categories. To address this issue, we propose an open-vocabulary multi-label action recognition method enhanced by the knowledge of large language models knowledge. This method extracts rich co-occurrence knowledge of action categories implicit in large language models and incorporates this co-occurrence knowledge into prompt learning of visual-language models, facilitating information transfer between base action classes and novel action classes to improve the recognition performance of novel action classes. We set up two ratios of base action classes to novel action classes in experiments, namely 3∶1 and 1∶1, represented as “75% seen” and “50% seen” respectively. Experimental results on AVA and MovieNet datasets show that compared with existing methods, when the base action classes are “75% seen”, our method improves the mAP metric for novel action recognition by 1.95% and 1.21% on AVA and MovieNet datasets, respectively. When faced with the more challenging scenario of “50% seen”, our method improves the mAP metric for novel action recognition by 2.59% and 1.06% on the two datasets, respectively.

Key words open-vocabulary action recognition; multi-label classification; prompt learning; large language model; CLIP model

摘 要 开放域多标签动作识别任务旨在对视频中训练阶段未见的人的多类动作进行识别.相较于传统动作识别,该任务更适应实际场景,具有广泛的应用前景.然而,开放域多标签动作识别具有很大的挑战性,需要将模型有效泛化到未见过的新动作类别.为了解决此问题,提出大语言模型知识引导的开放域多标签动作识别方法.该方法挖掘大语言模型蕴含的丰富的动作类别共现知识,并将共现知识嵌入视觉-语言模型的提示学习,实现基本动作类别（base action classes）与新动作类别（novel action classes）之间的信

收稿日期: 2024-06-17; **修回日期**: 2024-12-12

通信作者: 吴心筱（wuxinxiao@bit.edu.cn）

息传递,从而提升新类别的识别性能.在实验中将基本动作类别和新动作类别的比例设置为 3∶1 和 1∶1,分别表示为“75% 可见”和“50% 可见”.在 AVA 和 MovieNet 数据集上的实验结果表明,相较于现有方法,当基本动作类别为“75%”时,该方法在 2 个数据集的新动作类别识别指标 mAP 上分别提升了 1.95 个百分点和 1.21 个百分点;当面临基本动作类别为“50%”的更困难场景时,提出的方法在这 2 个数据集上新动作类别识别指标 mAP 上分别提升了 2.59 个百分点和 1.06 个百分点.

关键词 开放域动作识别;多标签分类;提示学习;大语言模型;CLIP 模型

中图法分类号 TP183

DOI: 10.7544/issn1000-1239.202440522 **CSTR:** 32373.14.issn1000-1239.202440522

视频中人的动作识别[1-5]是计算机视觉和人工智能领域备受关注的研究方向.目前多数研究局限于单一动作分类、有限动作集的动作识别.但在实际应用中,视频中的动作不受预定义标签的限制,随着场景和任务的动态变化,会出现新的多个动作类别,并可能在时间和空间上相互交叉和重叠.本文研究开放域多标签动作识别[6],旨在识别视频中出现的以及包括训练时未见的各种动作,在关系检测[7]、情绪检测[8]、视频理解[9-10]中拥有广泛的应用前景.

近年来,具备零样本(zero-shot)学习能力的视觉-语言模型(例如 CLIP[11])显著缩小了视觉和文本之间的语义差距,对齐了图像和文本的特征空间.因其出色的泛化能力,越来越多的研究者将视觉-语言模型应用于动作识别领域.目前大多数工作[2,12-13]通过学习提示或微调适配器等方式,匹配视频片段与文本描述以实现动作分类.针对有多个动作同时发生的复杂视频,Mondal 等人[6]则尝试将 CLIP 模型应用于多标签动作识别任务,编码每个动作类别标签并在视频特征中查询相应动作.虽然这些方法通过应用 CLIP,能将动作识别模型泛化至新的动作类别,但它们忽视了多类动作之间的共现关系,即不同动作在同一时间或同一场景中同时发生的情况.当动作发生在复杂场景时,动作的共现关系对多标签动作识别具有更加重要的作用.

本文提出了一种大语言模型(large language model, LLM)引导的开放域多标签动作识别方法.首先提问大语言模型以获取动作间的共现关系知识,并设计利用大语言模型知识的提示学习模块,然后通过结合动作共现关系的文本提示与视觉特征交互,在基本动作类别(base action classes)与新动作类别(novel action classes)间传递信息,进而实现开放域多标签动作识别.具体来说,首先根据动作交互方式将所有动作分为人与物交互、人与人交互、人自身状态 3 个大类.然后设计问题,如“please select {*num*} actions that may occur simultaneously with {a_i} in [*action_list*]$_j$”,a_i 表示在动作集合中的第 i 个动作,[*action_list*]$_j$ 为第 j 大类的所有动作,从大语言模型中获取动作共现关系知识.为了避免大语言模型的不稳定所带来的知识噪声,设计多轮询问的策略减少噪声以提升知识的准确性.最后,借助获得的知识,构建动作共现知识矩阵 $\boldsymbol{C} \in (0,1)^{N\times N}$,其中 N 是动作类别的数量,$C(i,j)=1$ 表示第 i 个动作与第 j 个动作之间有共现性,反之 $C(i,j)=0$ 则表示没有共现性.

在获得动作共现知识矩阵之后,本文设计了利用动作共现关系知识引导的提示学习模块,实现对训练时不可见的新动作类别的识别.具体来说,本文设计 2 种本文提示:一种是基于大语言模型知识的关系提示,通过动作共现知识矩阵使模型在基本动作类别中获得的信息传递到新动作类别上,提升模型的泛化能力;另一种是基于固定模板,形如“a photo of [CLASS]”的标准提示,通过标准提示来约束关系提示,维持训练稳定的同时使关系提示更符合 CLIP 模型的嵌入空间,提升模型的特征提取能力.

本文的主要贡献包括 3 个方面:

1)提出了一种新颖的大语言模型知识引导文本提示学习的开放域多标签动作识别方法,设计提问从大语言模型获取动作共现关系知识,并利用动作共现知识引导提示学习,将基本动作类别信息传递到新动作类别,提高了模型的泛化能力.

2)提出了结合共现关系知识的提示学习模型,通过对基于大语言模型知识的关系提示和基于固定模板的标准提示的学习,提升了模型的特征提取能力.

3)在 AVA 和 MovieNet 这 2 个数据集上进行了大量实验.实验结果表明,相较于现有方法,当基本动作类别为“75%”时,本文方法在 AVA 和 MovieNet 数据集上的新动作识别指标 mAP 分别提升了 1.95 个百分点和 1.21 个百分点;当基本动作类别为“50%”的更困难设置时,本文方法在 2 个数据集上的新动

作识别指标 mAP 分别提升了 2.59 个百分点和 1.06 个百分点.

1 相关工作

1.1 多标签动作识别

多标签动作识别[3,14-16]旨在识别视频中出现的多个动作. 早期工作[15-18]为每个动作类别训练一个二分类器. Dai 等人[15]和 Sozykin 等人[16]使用 3D 卷积神经网络处理视频的时序信息, 并使用多个二分类器集成神经网络作为分类器头来处理每个动作的分类. Tirupattur 等人[14]构建一个基于注意力的多标签动作依赖性(multi-label action dependency, MLAD)层中建模动作之间的共现依赖性和时序依赖性. Zhang 等人[19]为每个活动提取时空独立并且特定于活动的特征, 并学习不同类别之间的活动关联图, 以用于多标签识别.

与这些方法不同, 本文关注开放域的多标签动作识别. 训练时, 仅使用划分的基本动作类别, 在测试时还需要识别出训练未见过的新动作类别. 相较于大多数闭集分类的研究工作, 本文的工作更具有现实意义, 因为在日常生活中, 人们常常会面对更加错综复杂的情境和更加繁琐多样的动作.

1.2 开放域动作识别

视觉-语言预训练模型(如 CLIP)在开放域图像识别领域展现出了卓越的性能. 近年来, 研究人员开始将视觉-语言预训练模型引入开放域动作识别领域[2,12]. Wang 等人[2]利用 CLIP 丰富的模型知识, 通过添加时序融合层, 在视频中达到零样本动作识别能力. Ni 等人[20]提出帧间的注意力机制以获得视频帧之间的时序依赖信息, 并将信息融入 CLIP 模型, 使得视频特征和文本特征对齐. 然而, 上述工作解决的是单标签识别任务, 本文工作关注的是多标签任务, 更加贴合现实场景中的复杂动作识别需求.

Mondal 等人[6]使用 CLIP 构建多模态语义查询网络并将多模态融合信息作为 Transformer 解码器[21]中的查询部分, 生成用于多标签分类的表示. 本文使用从大语言模型提取的动作间共现知识融入提示学习, 有助于提升 CLIP 在开放域多标签动作识别上的识别能力.

2 开放域多标签动作识别方法

2.1 概 述

开放域多标签动作识别任务旨在识别包括训练时未见过的视频中多种动作类别. 具体而言, 所有动作类别被分为训练时可见的基本动作类别和训练时不可见的新动作类别. 模型在基本动作类别上进行训练, 但在测试时需要识别包含新动作类别在内的所有动作类别.

本文提出了一种基于大语言模型知识引导的开放域多标签动作识别方法. 首先, 通过设计问题向大语言模型询问动作与其他动作之间的共现关系, 以引导关系提示学习的训练, 其中关系提示通过联合共现动作与视觉特征交互, 确保提示学习能够泛化到新的动作类别. 同时, 引入标准提示来约束关系提示学习, 进一步增强模型的特征提取能力. 如图 1 所示, 本文提出的方法可以分为大语言模型知识获取模块和提示学习模块.

2.2 大语言模型知识获取模块

本文通过设计问题询问大语言模型获取动作共现知识. 首先根据动作的出现及动作交互方式不同, 本文将动作分为了 3 个大类: 人与物交互(例如“work on a computer”等)、人与人交互(例如“hand shake”等)、人自身状态(例如“walk”等). 基于这 3 个大类, 本文设置了“50% 可见”场景和“75% 可见”场景进行实验, 详细内容将在实验部分进行介绍. 在通过大语言模型获取动作共现知识时, 为了使得大语言模型的回答不会偏向某一交互方式, 本文按照大类针对每个动作分别向大语言模型提问, 获取每个动作的共现动作集合. 这一过程表示为

$$\begin{aligned}&\{A\}=\{\{A_1\},\{A_2\},\{A_3\}\},\\&r_i=Q(a_i,A_1)+Q(a_i,A_2)+Q(a_i,A_3),\end{aligned}\tag{1}$$

其中A表示数量为N的动作集合, A_k表示第k个动作大类($k=1,2,3$). Q代表提问的模板, 形如“please select $\{num\}$ actions that may occur simultaneously with $\{a_i\}$ in A_k”, a_i表示第i个动作, r_i表示大语言模型对第i个动作类别输出的共现动作集合, 图 2 展示了使用提问模板获取动作共现知识的示例.

基于从大语言模型中获取的共现关系知识$r=\{r_1, r_2, \cdots, r_N\}$, 本文提出了动作共现知识矩阵$\boldsymbol{C}\in(0,1)^{N\times N}$, 其中$N$是动作类别的数量, $C(i,j)=1$表示第i个动作与第j个动作之间有共现性, 反之$C(i,j)=0$则表示没有共现性. 所有动作类别包含基本动作类别和新动作类别. 对于新动作类别$C(i,j)$的值取决于动作j是否存在于i动作的共现关系知识r_i中. 对于基本动作类别, 本文从训练集真值中提取动作共现关系, 并对每个动作选出共现频率最高的k个动作. 为了与大语言模型输出的动作数保持一致, $k=10$.

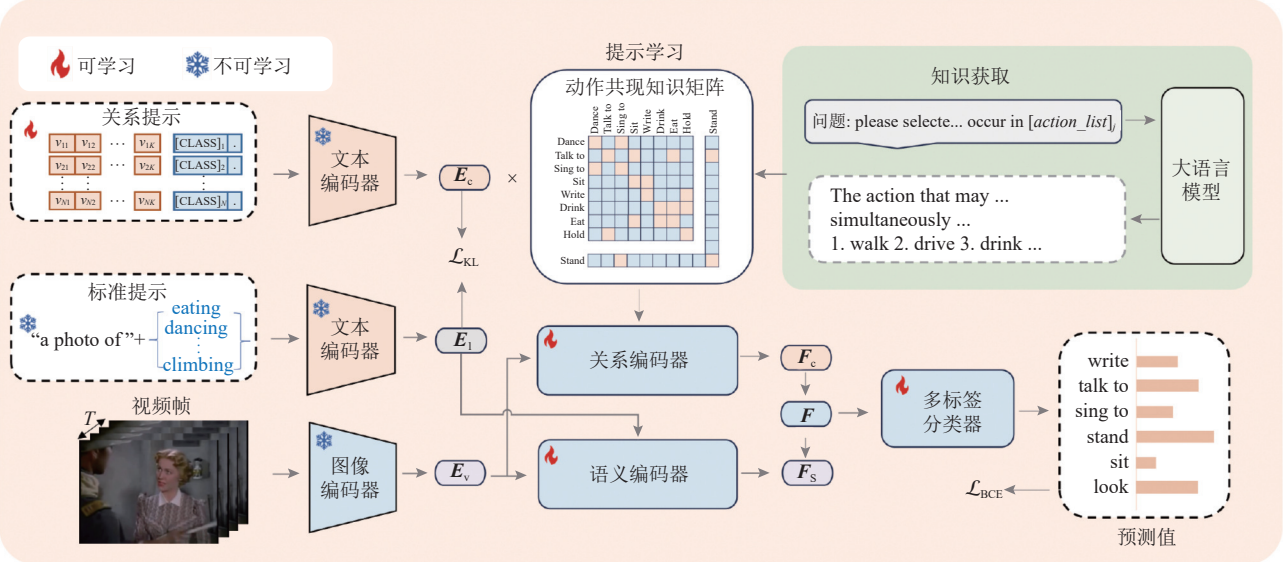


Fig. 1 Open-vocabulary multi-label action recognition architecture guided by LLM knowledge

图 1 大语言模型知识引导的开放域多标签动作识别架构

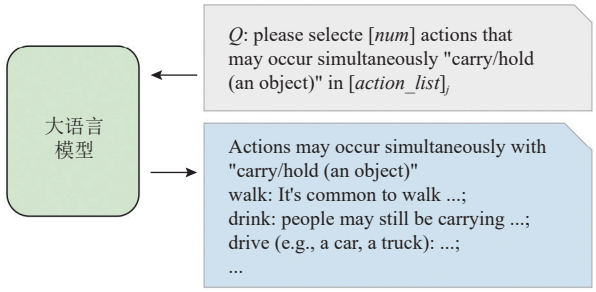


Fig. 2 Example of querying LLM

图 2 向大语言模型提问的示例

2.3 提示学习模块

本文提出了一个结合动作共现关系知识的提示学习模块以适应开放域多标签动作识别. 该模块引入基于知识共现矩阵的可学习关系提示和固定的标准提示作为查询，并通过基于注意力机制的解码器，与视觉特征交互以实现多标签分类. 在训练过程中，将动作共现矩阵作为基本动作类别和新动作类别之间的关联，实现从基本动作类别到新动作类别的信息传递. 因此，利用共现的基本动作特征提高了对新动作的识别能力. 更具体地说，提示学习模块包括CLIP 的文本和图像编码器、语义解码器、关系解码器和多标签分类器. 文本-图像编码器分别对文本提示和视频帧编码. 语义解码器将动作标签语义融合到视觉特征中，生成动作特定(action-specific)的语义特征. 关系解码器将动作共现知识融合到视觉特征中生成关系特征. 多标签分类器连接语义和关系特征进行预测，最后通过加入标准提示约束关系提示学习，提升模型特征提取能力.

1)图像-文本编码器. 本文利用 CLIP 的文本编码器 $\Psi_t(\cdot)$对可学习的关系提示和固定模板的标准提示编码. 本文所设置的关系提示为每一个动作设计数量为 K的学习标记(token)[$\boldsymbol{v}$]. 固定提示模板 $\boldsymbol{T}_1$形如“a photo of [CLASS]”，通过计算 KL 散度约束关系提示，使其更符合 CLIP 模型的嵌入空间，并维持模型训练的稳定性. 给定固定模板的标准提示 $\boldsymbol{T}_1$和可学习的关系提示 $\boldsymbol{T}_c$，动作特定的标签嵌入 $\boldsymbol{E}_1$和关系嵌入 $\boldsymbol{E}_c$计算为：

$$\begin{aligned}&\boldsymbol{t}_i=[\boldsymbol{v}]_1[\boldsymbol{v}]_2\cdots[\boldsymbol{v}]_K[\mathrm{CLASS}]_i,\\&\boldsymbol{T}_c=[\boldsymbol{t}_1,\boldsymbol{t}_2,\cdots,\boldsymbol{t}_N],\\&\boldsymbol{E}_1=\Psi_t(\boldsymbol{T}_1),\\&\boldsymbol{E}_c=\Psi_t(\boldsymbol{T}_c).\end{aligned}\tag{2}$$

给定视频 $\boldsymbol{V}\in\mathbb{R}^{T\times3\times H\times W}$，其中 $H\times W$表示空间维度，T表示帧数，本文利用 CLIP 图像编码器 $\Psi_v(\cdot)$将其编码为特征 $\boldsymbol{F}_v$：

$$\boldsymbol{F}_v=\Psi_v(V)\in\mathbb{R}^{T\times d},\tag{3}$$

其中 $d=512$表示嵌入维度. 另外，本文使用 KL 散度损失来约束关系提示和标准提示，表示为

$$\mathcal{L}_{\mathrm{KL}}=\phi_{\mathrm{KL}}(\boldsymbol{E}_1,\boldsymbol{E}_c).\tag{4}$$

2)语义解码器. 本文设计了语义解码器，以标签嵌入 $\boldsymbol{E}_1$作为查询(query)，视觉特征 $\boldsymbol{F}_v$作为键(key)和值(value)，利用注意力机制从动作帧中提取时空信息，获得用于分类的动作特定语义特征 $\boldsymbol{F}_s$. 语义解码器是一个掩码 Transformer 结构，包括多头自注意力(multi-head self-attention, MHSA)模块、多头交叉注意力(multi-head cross attention, MHCA)模块和前馈神经

网络(feed-forward neural networds, FFN). FFN 输出的语义特征$\boldsymbol{F}_{\mathrm{s}}$与标签嵌入在维度上保持一致，表示为

$$\begin{aligned}\hat{\boldsymbol{Q}}_{\mathrm{s}} &= \mathcal{M}_{\mathrm{MHSA}}(\boldsymbol{E}_{\mathrm{l}}),\\ \boldsymbol{Q}_{\mathrm{s}} &= \mathcal{M}_{\mathrm{MHCA}}(\hat{\boldsymbol{Q}}_{\mathrm{s}}, \boldsymbol{F}_{\mathrm{v}}),\\ \boldsymbol{F}_{\mathrm{s}} &= \mathcal{M}_{\mathrm{FFN}}(\boldsymbol{Q}_{\mathrm{s}}).\end{aligned} \tag{5}$$

3)关系解码器. 语义解码器有效地利用了 CLIP 的泛化能力，将视觉特征与其相应的动作类别标签对齐. 然而，在开放域动作识别任务中，我们认为还需要考虑到多标签动作依赖性，这种依赖性在共同出现的动作中至关重要. 为此，本文设计了一个关系解码器，它将动作共现知识矩阵$\boldsymbol{C}$和关系嵌入$\boldsymbol{E}_{\mathrm{c}}$融合并作为查询，将视觉特征$\boldsymbol{F}_{\mathrm{v}}$作为键和值，通过交叉注意力机制，将动作共现知识整合到视觉特征中，以生成关系特征$\boldsymbol{F}_{\mathrm{c}}$. 通过矩阵乘法将动作共现知识矩阵$\boldsymbol{C}$和关系嵌入$\boldsymbol{E}_{\mathrm{c}}$融合，使每个动作的嵌入与其共现动作的嵌入相加，形成一个复合查询. 这种方式能够学习共现动作之间的共享特征，并在反向传播的过程中共同更新共现动作提示的参数，实现这些动作之间的信息传递. 关系解码器包括一个多头交叉注意力模块和一个前馈网络，输出用于分类的关系特征$\boldsymbol{F}_{\mathrm{c}}$，表示为

$$\begin{aligned}\boldsymbol{Q}_{\mathrm{c}} &= \boldsymbol{C}\times\boldsymbol{E}_{\mathrm{c}},\\ \boldsymbol{Q}_{\mathrm{c}} &= \mathcal{M}_{\mathrm{MHCA}}(\boldsymbol{Q}_{\mathrm{c}}, \boldsymbol{F}_{\mathrm{v}}),\\ \boldsymbol{F}_{\mathrm{c}} &= \mathcal{M}_{\mathrm{FFN}}(\boldsymbol{Q}_{\mathrm{c}}).\end{aligned} \tag{6}$$

4)多标签分类器. 为了实现多标签动作分类，本文将语义特征$\boldsymbol{F}_{\mathrm{s}}$和关系特征$\boldsymbol{F}_{\mathrm{c}}$连接起来形成最终动作特定的特征$\boldsymbol{F}=[\boldsymbol{F}_{\mathrm{s}}|\boldsymbol{F}_{\mathrm{c}}]\in\mathbb{R}^{M\times 2d}$，其中$[\cdot|\cdot]$表示连接操作. 本文使用 sigmoid 函数$\sigma$和全连接(full connected, FC)层在训练过程中，对M个基本类别进行预测，采用二元交叉熵(binary cross entropy, BCE)损失和 KL 损失作为训练目标，表示为

$$\begin{aligned}&p_i = \sigma(\mathcal{M}_{\mathrm{FC}}(\boldsymbol{F}_i)), i\in\{1,2,\cdots,M\},\\ &\mathcal{L} = \sum_{i=1}^{M}\phi_{\mathrm{BCE}}(y_i, p_i) + \mathcal{L}_{\mathrm{KL}},\end{aligned} \tag{7}$$

其中，p_i表示被分类为第i个类别的预测概率，$y_i\in\{0,1\}$表示真实类别标签.

3 实验结果与分析

3.1 数据集

本文在 AVA[22] 和 MovieNet[23] 两个电影动作数据集上对模型进行了评估，其中的电影场景和动作分布与日常生活更接近. AVA 数据集包含了 430 个 15 min 的视频片段，标注了 80 种原始动作，如“sit”“stand”“walk”等. 本文使用 180 000 个视频片段进行训练，使用 50 000 个视频片段进行测试.

MovieNet 数据集提供了大量的电影动作片段，并为视频片段中的每个人物标注了边界框，用于视频理解和目标检测任务. 为了将数据集应用到多标签动作识别任务上，本文定义了 43 个动作类别，标注了大约 30 部电影，并使用了其中 15 000 个动作片段用于训练，3 000 个动作片段用于测试.

3.2 实验设置

本文将动作分为人与物交互、人与人交互以及人自身状态三大类. 根据训练时是否可见，将所有的动作分为基本动作类别和新动作类别，并将它们的比例分别设置为 3∶1 和 1∶1，表示为“75% 可见”和“50% 可见”. 在“75% 可见”场景下，每个大类中排名前 75% 的动作为基本动作类别，后 25% 的动作为新动作类别；在“50% 可见”场景下，前 50% 为基本动作类别，后 50% 为新动作类别. 通过这样划分基本动作类别和新动作类别是为了确保模型在训练过程中不会偏向某一个动作大类，并能够提取更为有效的动作共现知识. 这种处理方式有助于提升模型在开放域多标签动作识别任务中的泛化能力和准确性. 在训练时，模型使用基本动作类别训练. 在测试时，模型需要识别所有的动作类别，即计算在基本动作类别和新动作类别的识别性能. 本文使用 CLIP ViT/B-16[24] 中的视觉和文本编码器. 视频帧的分辨率设置为 224×224，模型使用 Adam[25] 优化器，将学习率设定为 0.000 003，总共进行 50 轮的训练. 本文采用的是与开放域多标签动作识别算法 MSQ-Net[6] 相同的评价指标，在每个数据批次计算 mAP[26] 和 $F1$[27] 分数，最终计算所有批次分数的平均值.

3.3 方法对比

本文与目前适用开放域多标签动作识别的多种方法进行对比，包括 Action-CLIP[2], BiAM[28], PS-ZSAR[29], DualCoOp[30], CoOp[31], CoCoOp[32], MSQ-Net[6].

表 1 展示了在 AVA 和 MovieNet 数据集上的对比实验结果. 从结果可以看出：本文方法在 AVA 和 MovieNet 数据集上的新动作类别都取得了最好的效果，这主要得益于动作共现知识引导的提示学习能够将基本动作类别学到的多标签动作共现信息有效传递给有共现关系的新动作类别. 此外，本文方法在基本动作类别上也取得了不错的结果，表明动作共现知识提示学习在促进多个类别间信息交互方面的有效性.

Table 1 Compared with the State-of-the-Art Methods on AVA and MovieNet Datasets

表 1 在 AVA 数据集和 MovieNet 数据集上与现有的方法作对比 %

类别划分	方法	AVA 数据集				MovieNet 数据集			
		50% 可见		75% 可见		50% 可见		75% 可见	
		mAP	*F*1	mAP	*F*1	mAP	*F*1	mAP	*F*1
基本动作	BiAM	42.24	25.67	36.20	27.14	35.18	23.96	30.87	23.51
	Action-CLIP	40.11	29.68	36.14	26.34	37.32	24.71	37.07	26.87
	PS-ZSAR	45.48	24.56	39.83	21.52	39.51	18.97	36.70	23.18
	DualCoOp	45.84	29.87	39.34	28.48	41.62	24.66	36.63	23.83
	CoOp	47.28	25.47	42.54	20.78	43.55	22.64	43.44	19.34
	CoCoOp	51.42	26.61	47.12	25.51	47.38	26.15	45.82	27.86
	MSQ-Net	46.87	26.99	46.77	25.77	51.22	25.84	41.03	24.03
	本文方法	**52.77**	**31.14**	**52.61**	**29.42**	**51.90**	**27.91**	**47.31**	**28.36**
新动作	BiAM	35.98	20.49	38.24	28.83	26.83	22.89	32.50	19.53
	Action-CLIP	33.37	23.55	29.67	24.54	33.94	21.27	34.39	18.81
	PS-ZSAR	35.84	18.36	38.61	23.89	30.45	18.38	32.66	20.41
	DualCoOp	35.49	25.51	39.01	29.13	30.07	21.82	33.89	19.25
	CoOp	36.43	21.52	40.44	22.73	37.25	16.61	40.73	23.71
	CoCoOp	38.68	22.51	41.79	26.37	39.47	20.62	41.64	21.74
	MSQ-Net	37.34	24.19	40.33	26.60	33.70	21.29	38.03	22.87
	本文方法	**41.27**	**26.80**	**43.74**	**30.75**	**40.53**	**23.89**	**42.85**	**24.11**

注：黑体数值表示最优结果.

3.4 消融实验

为验证本文方法中各个模块的有效性，以及在AVA数据集上验证关系提示和标准提示的有效性，设置了相应的消融实验. 结果如表 2 所示，可以看到，无论是去除关系提示还是去除标准提示，本文方法在基本动作类别和新动作类别上的性能都有所下降，这验证了关系提示和标准提示有效增强了模型在开放域下的动作识别能力.

Table 2 Effectiveness Verification Results of Relational Prompts and Standard Prompts on AVA Dataset

表 2 在 AVA 数据集上关系提示和标准提示的有效性验证结果 %

类别划分	C	S	50% 可见		75% 可见	
			mAP	*F*1	mAP	*F*1
基本动作	×	√	48.82	27.61	47.74	25.43
	√	×	51.77	30.16	48.39	25.17
	√	√	**52.77**	**31.14**	**52.61**	**29.42**
新动作	×	√	36.92	24.36	38.24	27.59
	√	×	37.32	25.94	40.11	27.07
	√	√	**41.27**	**26.80**	**43.74**	**30.75**

注：黑体数值表示最优结果. “C”代表关系提示，“S”代表标准提示. “√”表示含有相应模块，“×”表示去掉相应模块.

本文还在AVA数据集上验证关系提示和标准提示之间的KL损失的作用，结果如表 3 所示. 当去除KL损失之后在基本动作类别和新动作类别上的性能都有所下降，这验证了KL损失的有效性. 图 3 展示了本文方法在去掉关系提示的情况下在AVA数据集上不同动作类别的性能比较. 可以看到，本文方法利用关系提示改善了在大多数类别上的识别性能，体现了共现关系的有效性. 对于一些动作类别，比如“take from”和“sail boat”，本文方法并没有达到预期的效果，这可能因为大语言模型中的动作共现知识和数据集中电影片段中的动作共现分布存在差

Table 3 Effectiveness Verification Results of KL Divergence

表 3 KL 散度的有效性验证结果 %

类别划分	$\mathcal{L}_{KL}$	50% 可见		75% 可见	
		mAP	*F*1	mAP	*F*1
基本动作	×	52.71	30.87	52.07	26.58
	√	**52.77**	**31.14**	**52.61**	**29.42**
新动作	×	39.83	25.74	42.98	26.02
	√	**41.27**	**26.80**	**43.74**	**30.75**

注：黑体数值表示最优结果；$\mathcal{L}_{KL}$ 表示 KL 损失. “√”表示含有使用 KL 损失，“×”表示去掉 KL 损失.

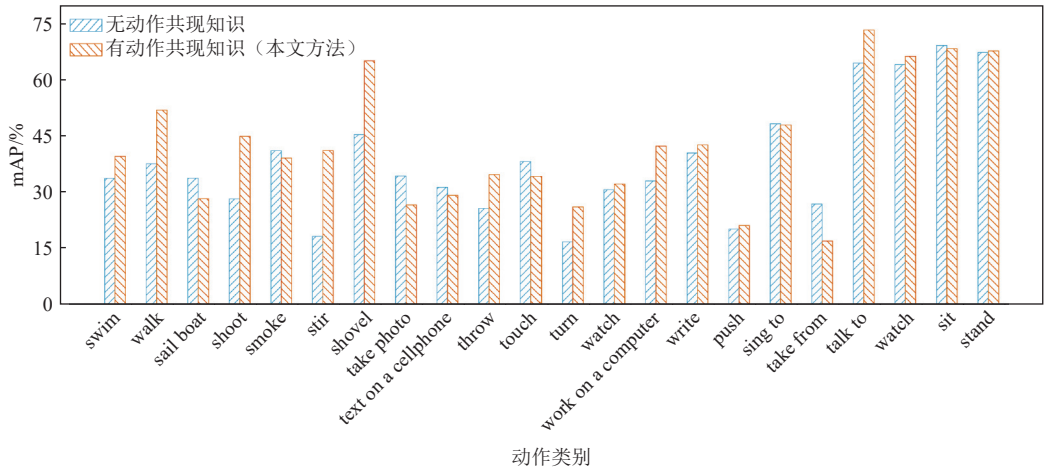


Fig. 3 Comparison of different categories with and without knowledge on AVA dataset

图 3 在 AVA 数据集上有无知识在不同类别上的比较

异导致的.

3.5 参数实验

为了探究关系提示中可学习标记数量K对于模型的影响，本文设计可学习标记数量K=8, 16, 32 的参数实验.

表 4 展示了在 2 个数据集上不同可学习标记数量的实验结果. 可以看到，在给定的可学习标记下，当关系提示中的K=16 时，模型能够达到相对较优的效果. 可能的原因是当学习的标记数量过少时，会限制模型学习到提示的表达能力，无法捕捉更细致的共现知识. 较多的可学习标记则增加了过拟合的风险，限制了模型的泛化能力.

Table 4 Experimental Results of Multiple Learned Marker Numbers on the Different Datasets

表 4 不同数据集上多种学习标记数量的实验结果 %

数据集	类别划分	K	50% 可见		75% 可见	
			mAP	$F1$	mAP	$F1$
AVA	基本动作	8	52.57	30.88	50.95	29.05
		16	**52.77**	**31.14**	52.61	29.42
		32	52.11	29.57	**54.43**	**30.74**
	新动作	8	38.56	23.44	40.14	26.87
		16	**41.27**	**26.80**	**43.74**	**30.75**
		32	40.52	26.51	39.16	27.30
MovieNet	基本动作	8	51.57	26.41	46.92	23.53
		16	51.90	27.91	47.31	28.36
		32	**52.34**	**28.09**	**47.53**	**28.66**
	新动作	8	**40.56**	22.38	40.14	23.57
		16	40.53	**23.89**	**42.85**	**24.11**
		32	38.71	21.06	41.89	23.90

注：黑体数值表示最优结果.

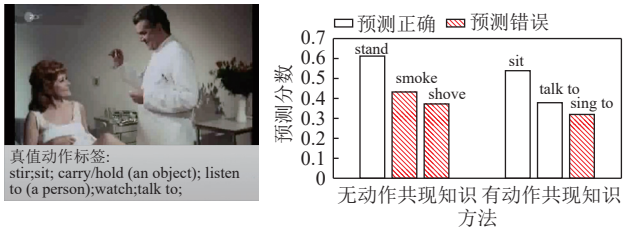
3.6 可视化分析

图 4 展示了没有引入共现知识方法和本文方法排名前三的动作类别预测结果. 正如我们所预料，通过添加动作共现知识，模型实现了更好的性能，如图 4(a)~(c)所示. 在图 4 (c)中，没有引入动作共现知识的方法错误地将人物动作分类为“take a photo”，而本文方法成功估计出了“watch”这一动作，这是因为本文方法从“shoot”这一动作传递相关信息，提高了“watch”的预测分数，结果展示了动作共现知识的优势.

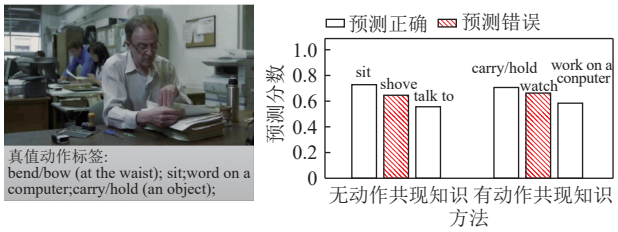
在图 4(d)中展示了本文方法预测的一个失败示例，其中打架场景中的“fight/hit”“jump/leap”动作被错误地分类为“push”“take from”“bend/bow”，我们推测这是因为从大语言模型获取的知识会在引导关系提示学习时将“fight/hit”更多地和手部动作关联，使模型更倾向预测“take from”和“push”等动作，误导了关系提示学习. 在未来的工作中，需要持续关注大语言模型动作共现知识与数据集动作分布存在差异以及知识去噪等问题.

4 总 结

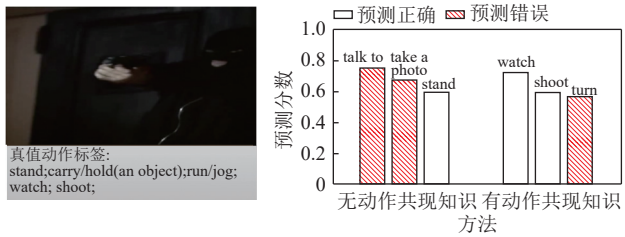
本文提出了大语言模型知识引导文本提示学习的开放域多标签动作识别方法. 本文通过设计提问从大语言模型获取动作共现知识，随后利用动作共现知识引导对 CLIP 模型的关系提示学习，在多标签开放域动作识别任务上取得了有效的进展. 在 AVA 和 MovieNet 数据集上大量的对比和消融实验验证了本文方法对动作识别的准确性和对各模块的有效性. 在未来工作中，我们将继续探究如何获得更加有效



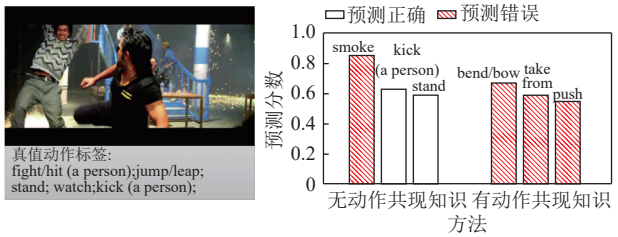
(a) 无动作共现知识方法与本文方法top-3准确率对比1:2



(b) 无动作共现知识方法与本文方法top-3准确率对比2:2



(c) 无动作共现知识方法与本文方法top-3准确率对比1:2



(d) 本文方法top-3预测全错

Fig. 4 Visualization of the action category prediction results with and without adding action co-occurrence knowledge

图 4 添加和未添加动作共现知识动作分类预测结果可视化

的动作关系知识以及如何处理知识中的噪声,探索不同的知识来源,提升模型对开放域动作的识别性能.

作者贡献声明:朱荣江负责设计与方法开发,监督整个研究进程,撰写论文及收集实验数据;石语珩负责方法分析与总结以及数据分析和结果解释,协助修订论文;杨硕实施实验与校对,协助修改论文;王子奕参与实验,协助数据处理;吴心筱监督整个实验过程,提供关键科学指导,全面修订和审阅论文.

参 考 文 献

[1] Wang Limin, Tong Zhan, Ji Bin, et al. TDN: Temporal difference networks for efficient action recognition[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1895−1904

[2] Wang Mengmeng, Xing Jiazheng, Liu Yong. Action-CLIP: A new paradigm for video action recognition[J]. arXiv preprint, arXiv: 2109.08472, 2021

[3] Munro J, Damen D. Multi-modal domain adaptation for fine-grained action recognition[C]//Proc of the 29th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 122−132

[4] Wang Xiang, Zhang Shiwei, Qing Zhiwu, et al. Molo: Motion-augmented long-short contrastive learning for few-shot action recognition[C]//Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 18011−18021

[5] Wang Chong, Wei Ziling, Chen Shuhui. Action identification without bounds on applications based on self-attention mechanism[J]. Journal of Computer Research and Development, 2022, 59(5): 1092−1104 (in Chinese)
(王冲, 魏子令, 陈曙晖. 基于自注意力机制的无边界应用动作识别方法[J]. 计算机研究与发展, 2022, 59(5): 1092−1104)

[6] Mondal A, Nag S, Prada J M, et al. Actor-agnostic multi-label action recognition with multi-modal query[C]//Proc of the 19th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2023: 784−794

[7] Xie Wentao, Ren Guanghui, Liu Si. Video relation detection with trajectory-aware multi-modal features[C]//Proc of the 28th ACM Int Conf on Multimedia. New York: ACM, 2020: 4590−4594

[8] Nandwani P, Verma R. A review on sentiment analysis and emotion detection from text[J]. Social Network Analysis and Mining, 2021, 11(1): 81

[9] Lin Ji, Gan Chuang, Han Song. TSM: Temporal shift module for efficient video understanding[C]//Proc of the 28th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2019: 7083−7093

[10] Wu Chaoyuan, Feichtenhofer C, Fan Haoqi, et al. Long-term feature banks for detailed video understanding[C]//Proc of the 28th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 284−293

[11] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[12] Gao Peng, Geng Shijie, Zhang Rrenrui, et al. Clip-adapter: Better vision-language models with feature adapters[J]. International Journal of Computer Vision, 2023, 12(6): 1−15

[13] Yang Taojiannan, Zhu Yi, Xie Yusheng, et al. AIM: Adapting image models for efficient video action recognition[J]. arXiv preprint, arXiv: 2302.03024, 2023

[14] Tirupattur P, Duarte K, Rawat Y S, et al. Modeling multi-label action dependencies for temporal action localization[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 1460−1470

[15] Dai Xiyang, Singh B, Ng Jy H, et al. TAN: Temporal aggregation network for dense multi-label action recognition[C]///Proc of the 19th IEEE Winter Conf on Applications of Computer Vision (WACV). Piscataway, NJ: IEEE, 2019: 151−160

[16] Sozykin K, Protasov S, Khan A, et al. Multi-label class-imbalanced action recognition in hockey videos via 3D convolutional neural networks[C]//Proc of the 19th IEEE/ACIS Int Conf on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. Piscataway, NJ: IEEE, 2018: 146−151

[17] Ji Shuiwang, Xu Wei, Yang Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 35(1): 221−231

[18] Yao Guangle, Lei Tao, Zhong Jiandan. A review of convolutional-neural-network-based action recognition[J]. Pattern Recognition Letters, 2019, 18(6): 14−22

[19] Zhang Yanyi, Li Xinyu, Marsic I. Multi-label activity recognition using activity-specific features and activity correlations[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 14625−14635

[20] Ni Bolin, Peng Houwen, Chen Minghao, et al. Expanding language-image pretrained models for general video recognition[C]//Proc of the 17th IEEE/CVF European Conf on Computer Vision. Berlin: Springer, 2022: 1−18

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. Berlin: Springer, 2017, 6000−6010

[22] Gu Chunhui, Sun Chen, Ross D A, et al. AVA: A video dataset of spatio-temporally localized atomic visual actions[C]//Proc of the 27th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 6047−6056

[23] Huang Qingqiu, Xiong Yu, Rao Anyi, et al. MovieNet: A holistic dataset for movie understanding[C]//Proc of the 16th IEEE/CVF European Conf on Computer Vision. Berlin: Springer, 2020: 709−727

[24] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010.11929, 2020

[25] Kingma D, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint, arXiv: 1412.6980, 2014

[26] Su Wanhua, Yuan Yan, Zhu Mu. A relationship between the average precision and the area under the roc curve[C]//Proc of the 5th Int Conf on the Theory of Information Retrieval. New York: ACM, 2015: 349−352

[27] Sasaki Y. The truth of the F-measure[J/OL]. Teach Tutor Mater, 2007[2024-09-19]. https://www.researchgate.net/publication/268185911_The_truth_of_the_f-measure

[28] Narayan S, Gupta A, Khan S, et al. Discriminative region-based multi-label zero-shot learning[C]//Proc of the 30th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 8731−8740

[29] Kerrigan A, Duarte K, Rawat Y, et al. Reformulating zero-shot action recognition for multi-label actions[C]//Proc of the 35th IEEE/CVF Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 25566−25577

[30] Sun Ximeng, Hu Ping, Saenko K. DualCoOP: Fast adaptation to multi-label recognition with limited annotations[C]//Proc of the 36th IEEE/CVF Annual Conf on Neural Information Processing Systems. Cambridge, MA: MIT, 2022: 30569−30582

[31] Zhou Kaiyang, Yang Jingkang, Loy C C, et al. Learning to prompt for vision-language models[J]. International Journal of Computer Vision, 2022, 130(9): 2337−2348

[32] Zhou Kaiyang, Yang Jingkang, Loy C C, et al. Conditional learning for vision-language models[C]//Proc of the 31st IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2022: 16816−16825

Zhu Rongjiang, born in 2001. Master candidate. His main research interest includes action recognition.

朱荣江, 2001 年生. 硕士研究生. 主要研究方向为动作识别.

Shi Yuheng, born in 2000. Master. His main research interest includes action recognition.

石语珩, 2000 年生. 硕士. 主要研究方向为动作识别.

Yang Shuo, born in 1992. PhD, associate professor. His main research interest includes video grounding.

杨　硕, 1992 年生. 博士, 副教授. 主要研究方向为时序视频定位.

Wang Ziyi, born in 2001. Master candidate. His main research interests include domain adaptation and domain generalization.

王子奕, 2001 年生. 硕士研究生. 主要研究方向为领域适应、领域泛化.

Wu Xinxiao, born in 1982. Professor, PhD supervisor. Member of CCF. Her main research interests include vision and language, machine learning, and video understanding.

吴心筱, 1982 年生. 教授, 博士生导师. CCF 会员. 主要研究方向为视觉与语言、机器学习、图像视频理解.