DOI:10.11784/tdxbz202403012

基于 CLIP 模型和知识数据库的零样本动作识别

侯永宏[1]，郑皓春[2]，高嘉俊[1]，任　懿[3]
（1. 天津大学电气自动化与信息工程学院，天津 300072；2. 天津大学未来技术学院，天津 300072；
3. 中国科学院软件研究所，北京 100190）

摘　要：零样本动作识别旨在从已知类别的动作样本数据中学习知识，并将其迁移到未知的动作类别上，从而实现对未知动作样本的识别和分类. 现有的零样本动作识别模型依赖有限的训练数据，可学习到的先验知识有限，难以将视觉特征准确地映射到语义标签上，是限制零样本学习性能提升的关键因素. 针对上述问题，本文提出了一种引入外部知识数据库和 CLIP 模型的零样本学习框架，利用多模态 CLIP 模型通过自监督对比学习方式积累的知识，来扩充零样本动作识别模型的先验知识. 同时，设计了时序编码器，以弥补 CLIP 模型时序建模能力的欠缺. 为了使模型学习到更丰富的语义特征，缩小视觉特征和语义标签之间的语义鸿沟，本文扩展了已知动作类别的语义标签，用更为详细的描述语句代替简单的文本标签，丰富了文本表示的语义信息；在此基础上，在模型外部构建了一个知识数据库，在不增加模型参数规模的条件下为模型提供额外的辅助信息，强化视觉特征与文本特征表示之间的关联关系. 最后，本文遵循零样本学习规范，对模型进行微调，使其适应零样本动作识别任务，提高了模型的泛化能力. 所提方法在 HMDB51 和 UCF101 两个主流数据集上进行了广泛实验，实验数据表明，该方法的识别性能相比目前的先进方法在上述两个数据集上分别提升了 3.8%和 2.3%，充分体现了所提方法的有效性.

关键词：零样本学习；动作识别；CLIP 模型；知识数据库
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：0493-2137（2025）01-0091-10

Zero-Shot Action Recognition Based on CLIP Model and Knowledge Database

Hou Yonghong[1]，Zheng Haochun[2]，Gao Jiajun[1]，Ren Yi[3]
（1. School of Electrical and Information Engineering，Tianjin University，Tianjin 300072，China；
2. School of Future Technology，Tianjin University，Tianjin 300072，China；
3. Institute of Software，Chinese Academy of Sciences，Beijing 100190，China）

Abstract：Zero-shot action recognition（ZSAR）aims to learn knowledge from seen action classes and apply it to unseen action classes，thereby achieving recognition and classification of unknown action samples. However，existing ZSAR models are limited by the amount of training data. This restricts their capability to learn prior knowledge and the accurate mapping of visual features with semantic labels. To address this issue，a ZSAR framework was proposed in this study by introducing an external knowledge database and using the contrastive language-image pre-training（CLIP）model. This framework utilized the knowledge acquired through self-supervised contrastive learning by the multimodal CLIP model to expand the prior knowledge of ZSAR. Moreover，a temporal encoder was designed to compensate for the lack of temporal modeling capability of the CLIP model. To enhance semantic features and bridge the gap between visual features and semantic labels，the semantic labels of seen action classes were extended. This involved replacing simple text labels with more detailed descriptive sentences to enrich the semantic information of text representations. On this basis，a knowledge database was constructed outside the model. This approach pro-

收稿日期：2024-03-11；**修回日期**：2024-04-28.
作者简介：侯永宏（1968—　　），男，博士，教授，houroy@tju.edu.cn.
通信作者：任　懿，englishhsl@126.com.
基金项目：国家自然科学基金资助项目（62102422）.
Supported by the National Natural Science Foundation of China（No. 62102422）.

vided additional information without increasing the model parameter scale and strengthens the association between the visual and text features. Finally, following the ZSAR protocol, the model was fine-tuned for the ZSAR task to improve its generalization ability. Furthermore, the proposed method was extensively experimented on two mainstream datasets: HMDB51 and UCF101. The experimental results demonstrate significant improvements of 3.8% and 2.3% on the above two datasets, respectively, compared with previous methods, validating the effectiveness of the proposed approach.

Keywords: zero-shot learning (ZSL); action recognition; contrastive language-image pre-training (CLIP) model; knowledge database

随着计算机视觉和人工智能技术的快速发展,视频理解技术取得了显著进步. 传统的有监督学习的方法需要依赖大量标准化数据进行训练,然而,随着数据集规模的不断扩大,逐一为视频样本进行人工标注显然不现实. 深度学习技术的蓬勃发展为解决数据标注难题提供了新的思路,零样本学习(zero-shot learning, ZSL)[1]应运而生,它旨在解决传统机器学习收集和标注数据困难的问题,是计算机视觉领域的研究热点之一. 零样本动作识别(zero-shot action recognition, ZSAR)[2]则是将零样本学习的理念和方法应用于动作识别中的一个研究方向,其目标是利用可见类标签的训练数据建立视频和文本标签之间的映射关系,从而实现对未知动作类别的识别.

目前主流的零样本动作识别方法主要是基于视觉特征和语义标签之间的映射关系进行识别[3],该方法通常包括视频特征提取、动作类别语义表示以及视频和语义表示之间的关系建立 3 个步骤. 在视频特征提取方面,早期工作主要采用手工构建特征的方法,如图像特征词典(bag of features, BoF)、HOG 等[4-5],然而手工标注数据的成本高昂,且难以收集不可见类的视频数据. 随着深度学习的广泛应用,多利用深度神经网络来提取视频特征,如C3D、I3D 等[6-8]. 在动作类别语义表示方面,早期工作主要采用手工定义的属性,如动作的持续时间、运动速度等[4-5,8],然而,手工定义属性同样受到主观性的影响,且难以描述复杂的动作类别. 近期多利用更详细的文本描述来表示动作类别,如词向量、句子嵌入等[9-12],进一步丰富了类别表示. 在建立视频和语义表示之间的关系方面,一部分工作[13]直接利用文本描述来代替视频,这种方法忽略了人体动作时序信息. 另一部分工作[14-15]认为视频和文本是同一视频的互补信息,通过额外添加辅助文本信息来增强视觉特征. 此外,当前可供使用的数据集中包含的标准化训练动作类别相对有限,导致模型能够学习到的可见类别相对较少[15-17]. 这限制了模型掌握的先验知识的规模,难以应对向众多新颖的不可见类动作的迁移与泛化.

针对上述问题,本文提出了一种基于多模态 CLIP 模型和知识数据库的零样本动作识别算法. 该算法将应用于图像分类领域的 CLIP 模型迁移到视频人体识别领域,借助 CLIP 模型海量的训练数据规模,扩充零样本动作识别模型的先验知识. 依照零样本的准则对 CLIP 模型在现有视频动作数据集上进行微调,以符合零样本学习的规范,同时使得 CLIP 模型适应视频人体动作识别任务. 重新构造类别语义标签,使用更为详细的语义描述替代简单的类别标签,提高了类别语义标签中有效信息的含量. 此外,本文还为模型引入了一个知识数据库,将更加详细的视频-文本对储存在模型外部,利用跨模态的检索信息进一步提高视频特征的表征能力,增强了模型对先验知识的进一步认知与应用能力. 实验结果表明,本文提出的算法在 HMDB51 和 UCF101 数据集上取得了显著成效,有效缓解了先验知识不足的问题.

1 相关工作

1.1 视频和类别表示

对于零样本动作识别的视频特征提取[18],早期的传统方法大多采用人工构建的特征描述符. 例如,Liu 等[4]构建了 BoF,但这种方法只关注了视频所涵盖的空间信息,忽略了时间维度的信息. Kodirov 等[5]进一步使用改进的密集轨迹(improved dense trajectories, IDT)算法,消除了背景的光流信息,抑制了背景运动所引起的噪声影响,从而获得更加聚焦于人体动作信息的特征,但是 IDT 算法引入了额外的计算开销,识别效果也不尽如人意. 总的来说,人工设计的方法受到主观注释的限制,而且在大规模数据集上进行手工注释是不现实的. 近年来,深度学习方法在视频特征提取方面取得了显著进展. 例如,Gan 等[10]和 Zhu 等[19]利用在大规模图像数据集 ImageNet 上预训练的 VGG、ResNet-200 等网络,提取视频帧级的特征,提高了特征提取的效率和精细程度. 然而,这些方法主要关注图像帧内的空间信息,而未能

有效处理动作的动态信息. 为了捕捉和理解视频的时序信息，一些研究者采用了三维卷积神经网络. 例如，Wang 等[20]采用了 C3D 网络，将二维卷积完全替换为三维卷积，将空间和时序信息整合在一个统一的网络结构下，从而提高了模型的性能. Brattoli 等[7]使用了 R(2 + 1)D 模型，将三维卷积核分解为二维的空间分量和一维的时间分量，从而使得模型更易于训练和优化，在参数量下降一半的情况下，提升了特征的精细程度. Chen 等[14]采用 TSM 模型[21]，通过时间位移来模拟三维网络建模，既保留了二维网络的处理效率，又提高了模型性能，确保了识别准确性. 这些方法能够同时处理视频的时空信息，但存在计算复杂度高、参数数量庞大等问题. 近期，视觉 Transformer 在视频特征提取方面取得了突破性进展. Lin 等[22]、Wang 等[16]以及 Ni 等[17]借助视觉 Transformer 的强大建模能力，提取视觉特征，使模型对视频时空信息的表征能力得到进一步提升.

在动作类别表示方面，传统的零样本动作识别方法采用人工设计的方式[3-5,8]来构建动作类别的标签. 但是这类方法难以客观地确定和划分属性，导致难以公平地定义众多平等的属性，同时还受到高昂人力成本的制约. 近年来随着深度学习的发展，基于词向量嵌入的动作类别表示方法[23]成为主流. Xu 等[9]采用词向量嵌入模型 Word2Vec，将动作类的名称作为语义标签，将其中离散的语言单词嵌入为连续表示的向量. Zhang 等[11]采用 Glove 模型进行动作类标签嵌入. 尽管基于词向量的方法简单高效，但存在一些局限性. 首先，动作类的名称不够精确，难以反映动作的细微差别. 其次，动作本身也难以仅根据其名称进行字面解释. 为了解决上述问题，Chen 等[14]提出了一种新的动作类别表示方法. 该方法将类别标签扩展为更为详细的概念，通过句子级别的概念扩充了对动作类的描述，从而增强了类别语义表示的丰富度和可辨别性. 然后，通过预训练的 BERT 模型提取类别的语义表示，将类概念编码作为语义特征，使得编码后的动作类语义表示更为精细化.

为了解决标签文本的不足的问题并利用大量网络数据，Wang 等[16]基于 Transformer 网络架构提出了 CLIP 模型. 该模型采集互联网上已有的海量的图像文本对，采用无监督学习方式训练图片和文本的配对关系，减少了对大量的标准化标注数据的依赖，在图像领域具有较好的零样本识别性能. 视频数据具有时序特征，而 CLIP 模型缺乏时序建模的能力，因此无法直接应用于视频领域. 为了使零样本动作识别任务能够充分利用 CLIP 模型以自监督对比学习训练的方式积累的知识，本文在 CLIP 模型的基础上添加了一个时序模块，赋予模型捕捉视频数据中时序特征的能力，从而实现其从图像领域向视频领域的迁移，并适应视频人体动作识别领域的零样本动作识别任务.

1.2　视频和类别语义关联

在零样本动作识别任务中，如何实现视觉特征空间和语义特征空间的有效关联[24]，并将此关联关系迁移到对不可见类视频动作识别中，是至关重要的. 目前基于深度学习的零样本动作识别方法主要有以下 3 种思路. 第 1 类方法是将视觉特征映射到语义特征空间中. Liu 等[4]提出视频的对象属性直接与类别属性相关联. Wang 等[20]和 Xu 等[9]将视觉特征映射到词向量嵌入构建的语义特征空间中，并在其中直接进行分类. 第 2 类方法是将视频和类别语义特征相关联，同时合成不可见的视觉特征. Zhang 等[12]引入了生成对抗网络(generative adversarial network，GAN)[25]，根据语义特征直接在视觉空间生成视觉特征，从而避免了特征空间中的显式投影. Lin 等[22]使用统一的 Transformer 模型进行跨模态表示，并通过学习语义一致的视觉表示来构建不可见的视频原型. Pu 等[26]则利用联合对比损失函数，分别来改善类间和类内特征的分布关系. 第 3 类方法是将视频和语义特征都映射到共享的公共空间中. Wang 等[20]提出了一个两阶段的模型，构建了一个潜在的嵌入空间. Gao 等[27]提出了一种双流框架，通过图卷积网络对视频和类别关系进行建模. Chen 等[14]引入了预训练的 BiT 模型[28]来捕捉视频帧内的物体信息，并将这些物体的名称扩展为句子级别的概念，编码为视频的语义特征，随后与视频视觉特征融合，用于增强视频的特征表示.

为了更深入地理解视频并增强其特征表示，促进模型利用语义知识引导视频与其相应的类别标签进行关联，本文提出了一种基于知识增强的新型零样本动作识别方法. 该方法的核心思想是在模型外部添加额外的知识数据库，并通过知识融合模块将知识信息融入到视觉特征和语义特征中，从而增强视觉空间和语义空间的关联.

2　方　法

2.1　算法框架

给定训练所用的可见类样本集合 $D^{\mathrm{s}}=\left\{\left(v_1^{\mathrm{s}},y_1^{\mathrm{s}}\right),\cdots,\left(v_{N_{\mathrm{s}}}^{\mathrm{s}},y_{N_{\mathrm{s}}}^{\mathrm{s}}\right)\right\}$，其中包含的视频序列片段 v_i^{s} 和对应的类别语义标签 y_i^{s} 对的数量为 N_{S}. 可见类标

签集合为 Y^{s}，$y_i^{\mathrm{s}} \in Y^{\mathrm{s}}$. 同理，定义测试推理阶段的不可见类样本集合 $D^{\mathrm{u}}=\left\{\left(v_1^{\mathrm{u}}, y_1^{\mathrm{u}}\right), \cdots,\left(v_{N_{\mathrm{u}}}^{\mathrm{u}}, y_{N_{\mathrm{u}}}^{\mathrm{u}}\right)\right\}$，不可见类标签集合 Y^{u} 与可见类标签不重叠，即 $Y^{\mathrm{s}} \cap Y^{\mathrm{u}}=\varnothing$. 零样本动作识别任务的目标便是基于 CLIP 模型，在训练阶段使用 D^{s} 对模型进行微调，学习可见类视频 v_i^{s} 和语义标签 y_i^{s} 之间的匹配关系，并将此匹配关系迁移泛化到测试阶段，实现对不可见类视频片段 v_i^{u} 的识别与分类.

本文所提的算法框架如图 1 所示. 对于一条视频 v_i，首先使用 CLIP 的视觉编码器 $G_{\mathrm{sp}}(\cdot)$ 逐帧处理其空间特征，生成视频帧的空间特征序列 $\{\boldsymbol{g}_{v_i}^{\mathrm{sp}}\}$. 随后，将视频帧的空间特征序列输入时序 Transformer 编码器 $G_{\mathrm{tem}}(\cdot)$，对时间维度进行处理与建模，得到融合了时间和空间信息的整体视频特征 $\boldsymbol{g}_{v_i}$. 对于视频 v_i 对应的类别语义标签 y_i，使用 CLIP 的文本编码器 $G_{\mathrm{text}}(\cdot)$ 对其进行特征提取，得到类别语义特征 $\boldsymbol{g}_{y_i}$. 随后，$\boldsymbol{g}_{v_i}$ 和 $\boldsymbol{g}_{y_i}$ 被投影到同一维度的公共特征嵌入空间中，得到视频的嵌入表示 $\boldsymbol{f}_{v_i}$ 和对应类别标签的嵌入表示 $\boldsymbol{f}_{y_i}$. 接下来，在向量数据库中，以 $\boldsymbol{f}_{v_i}$ 作为索引，通过 KNN 检索，获取其相近的特征 $\boldsymbol{f}_{v_i}^{\mathrm{m}}$，并返回 $\boldsymbol{f}_{v_i}^{\mathrm{m}}$ 对应的语义特征向量 $\boldsymbol{f}_{r_i}^{\mathrm{m}}$，用以增强 $\boldsymbol{f}_{v_i}$，得到 $\overline{\boldsymbol{f}_{v_i}}$. 在训练阶段 $\boldsymbol{f}_{v_i}^{\mathrm{s}}$ 和 $\overline{\boldsymbol{f}_{v_i}^{\mathrm{s}}}$ 被用于训练视频 v_i 和对应类别语义标签 y_i 之间的匹配关系.

在测试阶段，利用 $\boldsymbol{f}_{v_i}^{\mathrm{u}}$ 与不可见类标签集合 Y^{u} 内各个标签表示 $\boldsymbol{f}_{y_j}^{\mathrm{u}}$ 计算相似度，以寻找距离最近的特征，并返回相似度最高的标签 $\hat{y}_j^{\mathrm{u}}$ 作为预测的分类结果.

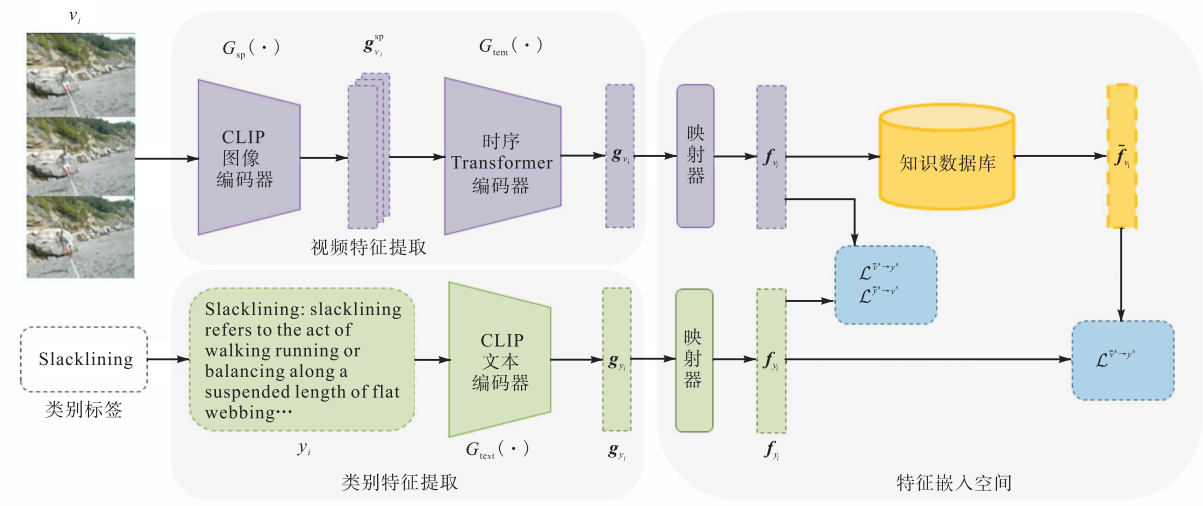

图 1 基于 CLIP 和知识数据库的动作识别框架

Fig.1 Structure of action recognition based on CLIP and knowledge database

2.2 多模态特征提取

2.2.1 视频特征提取

在本文设计的算法网络结构中，采用了两个 Transformer 编码器处理视频来提取特征. 第 1 个 Transformer 网络为 CLIP 模型的图像编码器 $G_{\mathrm{sp}}(\cdot)$，该编码器经过了预训练，用于处理各个视频帧的空间特征.

对于给定一段视频 v_i，首先将其采样为 T 帧的视频序列 $\{v_{i,1}, v_{i,2}, \cdots, v_{i,T}\}$，其次，对于视频序列中的一个视频帧 $v_{i,j}$，先进行分块处理，对各个块加上空间位置标记后，再利用 $G_{\mathrm{sp}}(\cdot)$ 提取这一帧的空间特征，即

$$g_{v_{i,j}}^{\mathrm{sp}}=G_{\mathrm{sp}}\left(v_{i,j}\right) \tag{1}$$

T 个视频帧均通过同样的处理，得到视频 v_i 的空间特征序列 $\boldsymbol{g}_{v_i}^{\mathrm{sp}}$ 为

$$\boldsymbol{g}_{v_i}^{\mathrm{sp}}=\left\{g_{v_{i,1}}^{\mathrm{sp}}, g_{v_{i,2}}^{\mathrm{sp}}, \cdots, g_{v_{i,T}}^{\mathrm{sp}}\right\} \tag{2}$$

接下来，空间特征序列 $\boldsymbol{g}_{v_i}^{\mathrm{sp}}$ 输入第 2 个时序 Transformer 网络 $G_{\mathrm{tem}}(\cdot)$，以捕捉视频帧序列间的时序信息.

具体而言，首先按照时间顺序为 $\boldsymbol{g}_{v_i}^{\mathrm{sp}}$ 中的每一条 $g_{v_{i,j}}^{\mathrm{sp}}$ 加上时序标记，随后利用 $G_{\mathrm{tem}}(\cdot)$ 对视频帧序列的时间维度进行建模得

$$\boldsymbol{g}_{v_i}=G_{\mathrm{tem}}\left(\boldsymbol{g}_{v_i}^{\mathrm{sp}}\right) \tag{3}$$

然后，$\boldsymbol{g}_{v_i}$ 经过全连接层，被映射到公共的嵌入空间中，即

$$\boldsymbol{f}_{v_i}=W^v \boldsymbol{g}_{v_i}+b^v \tag{4}$$

式中：$\boldsymbol{f}_{v_i} \in \mathbf{R}^K$ 为视频 v_i 的嵌入表示；K 为公共嵌入空间的特征维度；W 为权重.

2.2.2　类别标签特征提取

在处理类别语义标签特征之前，先对动作类的语义标签进行扩容与增强. 具体来说，在传统的基于 CLIP 模型的方法中，通常使用“标签 + 提示”[29-30]的形式构造类别语义标签，常见的提示一般为“a photo of {}”、“this is a {} photo”等形式，随后用填空的方式在空格中填入原始类别的标签名称. 由此可见，作为类别语义标签的一句话中，有效部分仅为原始类别标签的单词，且原始类别标签一般只有一两个单词[31]，这样导致类别语义标签中有效信息含量过低. 同时，单纯使用原始类别标签的单词不足以体现一个动作类别应该包含的信息量，而且对于一些动作来说，仅由标签的字面意思是难理解的. 为了提高类别语义标签的信息量和描述性，笔者依据目前零样本动作识别的常见方法，对模型所使用的动作类的语义标签进行了改造. 具体来说，首先是增强标签描述性，对标签进行语义分析，并加入一些描述性的词语，例如动作发生的场景、动作的主体、动作的工具等；其次优化标签格式，将标签格式调整为更易于理解和识别的形式，例如使用句子或短语的形式来描述动作. 图 2 展示了经过细化和增强的标签与传统标签的对比情况.

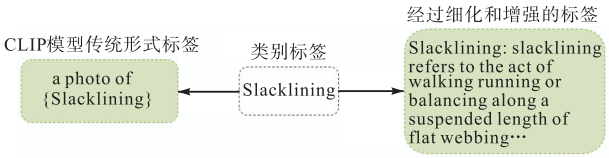

图 2　经过细化和增强的标签与传统标签的对比

Fig.2　Comparison between refined and enhanced labels and traditional labels

通过对比可见，使用比类别标签更为详细的描述语句替代了单纯的类别标签，不仅提高了类别语义标签中有效信息的含量，还使得对于动作类的描述更为细致，使模型能够区分更多的细节信息，从而更深入地学习视频动作，也有效地提高了模型对知识的泛化能力.

经过增强的动作类语义标签 y_i 以单词序列 $y_i=\{y_{i,1},y_{i,2},\cdots,y_{i,l_i}\}$ 的形式呈现，其中 l_i 表示 y_i 单词序列的长度. y_i 被输入到 CLIP 的文本编码器 $G_{\text{text}}(\cdot)$ 中进行编码，取编码器输出的句子级别的特征 $\boldsymbol{g}_{y_i}$ 作为类别语义标签的特征，即

$$\boldsymbol{g}_{y_i}=G_{\text{text}}(y_i) \tag{5}$$

随后，$\boldsymbol{g}_{y_i}$ 经过全连接层被映射到公共的嵌入空间之中即

$$\boldsymbol{f}_{y_i}=W^y\boldsymbol{g}_{y_i}+b^y \tag{6}$$

式中 $\boldsymbol{f}_{y_i}\in\mathbf{R}^K$ 为类别语义标签 y_i 的嵌入表示.

2.3　知识数据库的构建

为了进一步提高视频特征的精细程度，本文引入文本语义信息来增强视频的嵌入表示. 传统的做法是将文本语义信息直接编码到模型参数之中，然而，将过多细粒度的文本知识直接编码到模型参数之中会导致模型参数的数量大幅度增加，降低模型的训练效率. 模型直接学习文本语义信息会导致模型对训练数据过拟合，使模型泛化能力受限. 为了解决上述问题，本文构建了一个知识数据库，将文本语义信息存储在模型外部. 在训练过程中，模型首先提取视频的视觉特征，然后根据视频的视觉特征检索相关的知识向量，最后将视觉特征和知识向量融合在一起，形成最终的视频嵌入表示. 这种融合机制可以帮助模型学习到更丰富的语义信息，从而更好地区分不同动作类别.

在遵守零样本动作识别统一准则的前提下[7,22,26]，知识向量数据库以随机采样 Kinetics 数据集[32-33]中的视频-文本对的形式构建，其中的视频与文本描述分别以经过各自特征提取器，并以特征向量的形式存储和使用. 在训练阶段，对于视频 v_i，使用其嵌入表示 $\boldsymbol{f}_{v_i}$ 作为查询，在向量库中利用 KNN 检索得到其最邻近的向量 $\boldsymbol{f}_{v_i}^{\text{m}}$，并返回与 $\boldsymbol{f}_{v_i}^{\text{m}}$ 配对的语义特征向量，即

$$\boldsymbol{f}_{r_i}^{\text{m}}=\text{KNN}_{\text{text}}^{v\to v}\left(\boldsymbol{f}_{v_i}^{\text{m}},M\right) \tag{7}$$

随后，使用均衡合并相同通道的形式完善原始的 $\boldsymbol{f}_{v_i}$，得到 $\overline{\boldsymbol{f}_{v_i}}$ 为

$$\overline{\boldsymbol{f}_{v_i}}=\frac{\boldsymbol{f}_{v_i}+\boldsymbol{f}_{r_i}^{\text{m}}}{2} \tag{8}$$

将 $\overline{\boldsymbol{f}_{v_i}}$ 与 $\boldsymbol{f}_{v_i}$ 一起在训练阶段用于构造训练的损失函数.

3　实　验

3.1　训练与测试

在训练阶段，对于一个训练批次内可见类样本的视频——类别语义标签对 $\left(\boldsymbol{v}^{\text{s}},\boldsymbol{y}^{\text{s}}\right)$，首先分别提取在公共嵌入空间中的特征嵌入表示 $\boldsymbol{f}_v^{\text{s}}$ 和 $\boldsymbol{f}_y^{\text{s}}$，其中每一行代表一条具体可见类样本的 $\boldsymbol{f}_{v_i}^{\text{s}}$ 和 $\boldsymbol{f}_{y_i}^{\text{s}}$. 对于 $\boldsymbol{f}_{v_i}^{\text{s}}$，还利用辅助知识数据库构建了其增强的嵌入表示 $\overline{\boldsymbol{f}_{v_i}}$. 随后，使用 KL 散度构造训练损失函数. 在设计损失

函数时,为了衡量一个训练批次内视频与类别语义标签对的相似性,基于余弦距离定义了视频与类别语义标签对称相似度分别为

$$\mathrm{sim}\left(\boldsymbol{v}^{\mathrm{s}}, y_i^{\mathrm{s}}\right)=\frac{\boldsymbol{f}_v^{\mathrm{s}}\left[\boldsymbol{f}_{y_i}^{\mathrm{s}}\right]^{\mathrm{T}}}{\left\|\boldsymbol{f}_v^{\mathrm{s}}\right\|\left\|\boldsymbol{f}_{y_i}^{\mathrm{s}}\right\|} \tag{9}$$

$$\mathrm{sim}\left(\boldsymbol{y}^{\mathrm{s}}, v_i^{\mathrm{s}}\right)=\frac{\boldsymbol{f}_y^{\mathrm{s}}\left[\boldsymbol{f}_{v_i}^{\mathrm{s}}\right]^{\mathrm{T}}}{\left\|\boldsymbol{f}_y^{\mathrm{s}}\right\|\left\|\boldsymbol{f}_{v_i}^{\mathrm{s}}\right\|} \tag{10}$$

计算得到的 $\mathrm{sim}\left(\boldsymbol{v}^{\mathrm{s}}, y_i^{\mathrm{s}}\right)$ 为一个批次内视频样本 $\boldsymbol{v}^{\mathrm{s}}$ 与一条类别语义标签 y_i^{s} 的相似度矩阵,进而,基于这种相似度计算的方法,经 softmax 计算从视频到类别语义标签的相似度分数为

$$p_i^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}\left(\boldsymbol{v}^{\mathrm{s}}\right)=\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{v}^{\mathrm{s}}, y_i^{\mathrm{s}}\right)/\tau\right)}{\sum\limits_{j\in N_{\mathrm{B}}}\exp\left(\mathrm{sim}\left(\boldsymbol{v}^{\mathrm{s}}, y_j^{\mathrm{s}}\right)/\tau\right)} \tag{11}$$

式中:τ 为温度系数;N_{B} 为训练批次的大小.

同理,计算从类别语义标签到视频的相似度分数为

$$p_i^{\boldsymbol{y}^{\mathrm{s}}\to v^{\mathrm{s}}}\left(\boldsymbol{y}^{\mathrm{s}}\right)=\frac{\exp\left(\mathrm{sim}\left(\boldsymbol{y}^{\mathrm{s}}, v_i^{\mathrm{s}}\right)/\tau\right)}{\sum\limits_{j\in N_{\mathrm{B}}}\exp\left(\mathrm{sim}\left(\boldsymbol{y}^{\mathrm{s}}, v_j^{\mathrm{s}}\right)/\tau\right)} \tag{12}$$

$q_i^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}\left(\boldsymbol{v}^{\mathrm{s}}\right)$ 表示从视频到类别语义标签分类的真实值,即

$$q_i^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}\left(\boldsymbol{v}^{\mathrm{s}}\right)=\begin{cases}1 & j=i\\ 0 & j\neq i\end{cases} \tag{13}$$

同理,构造类别语义标签到视频分类的真实值为

$$q_i^{\boldsymbol{y}^{\mathrm{s}}\to v^{\mathrm{s}}}\left(\boldsymbol{y}^{\mathrm{s}}\right)=\begin{cases}1 & j=i\\ 0 & j\neq i\end{cases} \tag{14}$$

对于从视频到类别语义标签分类,以 KL 散度构造损失函数项为

$$\mathcal{L}^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}=\mathrm{KL}\left(\boldsymbol{p}^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}\left(\boldsymbol{v}^{\mathrm{s}}\right), \boldsymbol{q}^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}\left(\boldsymbol{v}^{\mathrm{s}}\right)\right) \tag{15}$$

类似地,分别利用 $\boldsymbol{f}_y^{\mathrm{s}}$ 与 $\boldsymbol{f}_v^{\mathrm{s}}$ 以及 $\overline{\boldsymbol{f}_v^{\mathrm{s}}}$ 与 $\boldsymbol{f}_y^{\mathrm{s}}$ 构造损失函数项,最终得到整体的训练损失函数为

$$\mathcal{L}=\frac{1}{3}\left(\mathcal{L}^{\boldsymbol{v}^{\mathrm{s}}\to y^{\mathrm{s}}}+\mathcal{L}^{\boldsymbol{y}^{\mathrm{s}}\to v^{\mathrm{s}}}+\mathcal{L}^{\overline{\boldsymbol{v}}^{\mathrm{s}}\to y^{\mathrm{s}}}\right) \tag{16}$$

在测试阶段,利用 $\boldsymbol{f}_{v_i}^{\mathrm{u}}$ 与不可见类标签集合 Y^{u} 内各个标签表示 $\boldsymbol{f}_{y_j}^{\mathrm{u}}$ 计算相似度,并返回相似度最高的标签 $\hat{y}_j^{\mathrm{u}}$ 作为预测的分类结果.

3.2 数据集与评价指标

本文的实验在 3 个公开数据集上进行,分别为 Kinetics[32-33]、UCF101[34]以及 HMDB51[35]. 这 3 个数据集是零样本动作识别领域最常用的性能评估数据集,具有较高的代表性和可靠性. Kinetics 数据集包含超过 40 万条视频样本,涵盖了 700 个动作类,是目前规模最大的动作识别数据集;HMDB51 数据集包含 6 766 个视频样本,涵盖了 51 个动作类别;UCF101 数据集包含 13 320 个视频样本,涵盖了 101 个动作类别.

本文实验采用 Top-1 识别准确率($A_{\text{Top-1}}$)作为性能评价指标. 对于每个不可见类的测试视频样本,如果网络输出的分类预测结果标签与其真实标签一致,则定义为分类正确,否则分类错误. 而 Top-1 识别准确率是指对于所有视频样本分类相似度分数最高的动作类标签与其真实标签一致的比例,计算公式为

$$A_{\text{Top-1}}=\frac{N_{\text{correct}}}{N_{\mathrm{u}}} \tag{17}$$

式中:N_{u} 为测试过程中所有不可见类视频样本的总数;N_{correct} 表示分类正确的视频样本数量. 对于每个测试视频样本,模型会输出一个包含所有类别预测概率的向量,将统计所有测试视频样本中预测类别与真实类别一致的样本数量预测概率最高的类别作为预测类别,将统计得到的分类正确的样本数量除以总测试视频样本数量,得到准确率.

本文在上述 3 个数据集的基准上采用与文献[36]中相同的划分方式——50% 50% 划分. 具体来说,整个数据集被随机划分成多个独立的分割片段,每个分割片段中的 50% 动作类别用于训练,而剩余的 50% 则在推理阶段作为不可见类用于测试. 这种划分方法可以确保训练集和测试集中的动作类别完全不重叠,从而更客观地评估模型的性能. 依据零样本动作识别领域目前唯一的准则[7,22,26],从 Kinetics-600 数据集中挑选了与 UCF101 ∪ HMDB51 不重叠的 564 个类别,用于在训练阶段对 CLIP 模型进行微调. 在测试阶段,直接使用 UCF101 和 HMDB51 数据集经过 50% 50% 划分得到的后 50% 类进行识别测试. 对于构建知识数据库,则是在 564 个类别中随机采样了 16 920 条视频-文本对,并将它们进行编码,以特征向量的形式进行存储.

3.3 实验参数设置

本文的实验均在 Ubuntu18.04 操作系统下设计实现. 程序设计语言采用广泛应用的高级编程语言 Python,本文所提算法的实现与验证均基于深度学习框架 PyTorch.

输入的每条视频样本都被采样为 $T\times H\times W$ 的视频帧片段,其中采样帧数 T 设置为 8,每帧图片的长 H 和宽 W 均被裁切为 224 像素. 公共特征嵌入空间的维度 K 设置为 512. 模型的视觉编码器是基于

CLIP 模型的 ViT-B/32 网络，由 12 个编码层组成，输入的视频帧被切割为 32 个图片块. 模型的时间视觉编码器采用了包含 6 层编码器的 Transformer 网络，模型的文本编码器是基于 CLIP 的 12 层 Transformer 编码器. 在对模型进行微调时，使用了 Kinetics 数据集，初始学习率设置为 5×10^{-6}，批次大小设置为 128，对学习率进行调整，伴随着预热与余弦退火. 采用 ADamW 优化器. 在训练阶段，本文模型一共训练 20 个循环(epoch)，权重衰减为 0.2，学习率每 10 个迭代轮数以 0.1 的衰减率进行调整，并随着训练更新向量数据库. 在测试阶段，分别在 UCF101 和 HMDB51 数据集上各随机选取了 3 个不同的划分进行识别测试，分别包含 50 个和 26 个动作类进行分类识别测试，以平均的 Top-1 识别准确率(%)±标准差的形式展现实验的结果.

3.4　实验结果

表 1 展示了所提方法在 UCF101 和 HMDB51 两个公开数据集上与当前多个先进的零样本动作识别算法的对比，在该表中，一些研究工作[19,37]采用 Fisher 向量(Fisher vector，FV)用于视频表示，另一些代表性的方法[14]则使用物体的特征(Obj)进行视频表示，还有一些研究工作[7-8,25,38]则是采用三维卷积神经网络提取视觉特征. 大多数研究采用手工设计的属性或词向量嵌入来表示动作类的语义信息，而 ER 模型[14]则是利用类别概念表示动作类别. 实验结果表明，本文所提方法的性能超越了现有的最先进的零样本动作识别算法，分别在 UCF101 和 HMDB51 数据集上获得了 3.8% 和 2.3% 的性能提升，充分体现了本文算法的有效性. 与基于有限少量数据构建的模型[7,14,26]相比，本文所提算法模型的优异表现验证了引入多模态 CLIP 模型以及在零样本学习的规范下扩大训练可见先验知识规模的必要性. 与同样基于 CLIP 模型的方法[15-17]相比，本文方法在微调阶段规避了重叠类别问题，并改进了类别语义标签的构造形式，同时构建了额外的知识数据库，在遵循零样本学习规范的前提下取得了优异的性能.

表 1　不同方法在两个数据集上识别准确率对比

Tab.1　Comparison of recongnition accuracy among different methods on two datasets

方法	视频特征	类别特征	在 UCF101 数据集上的识别准确率/%	在 HMDB51 数据集上的识别准确率/%
ESZSL[37]	FV	W	15.0 ± 1.3	18.5 ± 2.0
ASR[8]	C3D	W	24.4 ± 1.0	21.8 ± 0.9
UR[19]	FV	W	17.5 ± 1.6	24.4 ± 1.6
TRAN[38]	C3D	W	19.0 ± 2.3	19.5 ± 4.2
CEWGAN[8]	I3D	W	26.9 ± 2.8	30.2 ± 2.7
TS-GCN[27]	GCN	W	34.2 ± 3.1	23.2 ± 3.0
E2E[7]	R(2 + 1)D	W	48.0	32.7
ER[14]	S + Obj	ED	51.8 ± 2.9	35.3 ± 4.6
AURL[25]	R(2 + 1)D	W	58.0	39.0
SPOT[39]	Obj	W	40.9	35.9
ActionCLIP[16]	CLIP	W	69.6	50.1
本文方法	CLIP	ED	73.4 ± 2.1	52.4 ± 4.1

3.5　消融实验

为了探索微调模型的有效性以及重新构造语义类别标签和构建知识数据库对模型的影响，本节在 50%/50% 划分的评价规范下，分别在 UCF101 和 HMDB51 数据集上进行了消融实验.

3.5.1　验证微调模型的有效性

为了验证在 CLIP 模型上针对于零样本动作识别任务进行微调的有效性，本文对微调前后模型在 UCF101 和 HMDB51 两个数据集上的识别准确率进行了比较，结果如表 2 所示. 表中分别是直接使用基础的经过预训练的 CLIP 模型的识别准确率，以及剔除了可能存在重叠的类别后，在 Kinetics-600 的子集上继续对模型进行微调后的识别准确率. 通过比较可以看出，在对模型进行针对特定视频人体动作识别与零样本动作识别任务的微调后，测试阶段在两个数据集上的识别准确率分别提升了 32.5% 和 22.7%，相比基础的 CLIP 模型与时间视觉编码器的组合得到了显著的提升. 这证明了针对特定任务进行模型微调的有效性，能够使先验知识更好地符合特定下游任务的需求，提高了模型的泛化能力与识别性能.

表 2　微调前后模型识别准确率对比

Tab.2　Comparison of model recognition accuracy before and after fine-tuning

模型	在 UCF101 数据集上的识别准确率/%	在 HMDB51 数据集上的识别准确率/%
基础 CLIP 模型 + 时间视觉编码器	34.9 ± 3.0	26.3 ± 8.8
基础 CLIP 模型 + 时间视觉编码器 + Kinetics-564 微调	67.4 ± 0.9	49.0 ± 3.7

3.5.2　验证重新构造类别语义标签的有效性

为了提高类别语义标签的有效信息含量和表达能力，本文所提出的模型在提取类别语义信息之前，首先对动作类的语义标签进行了扩容与增强. 具体来说，将传统的简短文本标签替换为更详细的描述性语句，以丰富文本表示中的语义信息. 为了验证重新构造的经过增强的类别语义标签的有效性，本文对使用传统 CLIP 语义标签的模型以及使用改造后语义标签的模型的性能进行了比较，结果如表 3 所示. 表中分别是在微调与测试阶段使用传统的“标签+提

示”形式类别标签和使用详细的描述语句作为类别标签的识别准确率. 通过比较可见,在微调与测试阶段使用详细的类别概念描述语句作为语义标签后,模型在 UCF101 和 HMDB51 数据集上分别得到了 5.1% 和 2.0% 的性能提升. 识别准确率的提高证明使用改造后语义标签的模型可以明显提高识别性能.

表 3 使用不同语义标签的识别准确率对比

Tab.3 Comparison of recongnition accuracy using different categories of semantic labels

语义标签	在 UCF101 数据集上的识别准确率/%	在 HMDB51 数据集上的识别准确率/%
标签 + 提示	67.4 ± 0.9	49.0 ± 3.7
类别概念描述语句	72.5 ± 1.9	51.0 ± 4.9

3.5.3 验证构建知识数据库的有效性

为了进一步增强视频特征嵌入的表示能力,同时避免给模型引入繁重的训练参数负担,本文提出的算法在模型外部构建了一个知识数据库. 通过同模态的视觉特征作为查询条件,在知识库中匹配相近的视觉特征,并返回跨模态的语义特征,用于增强原视觉特征. 为了验证构建知识数据库的有效性,本文对有无使用外部知识数据库的模型的识别性能进行了比较,结果如表 4 所示. 表中分别是在改进类别标签基础上是否使用知识数据库的模型识别准确率. 通过比较可以看出,在使用知识数据库进行特征增强后,模型在 UCF101 和 HMDB51 数据集上的识别准确率分别提高了 0.9% 和 1.4%,这表明构建知识数据库能够有效地增强视频特征嵌入的表示,从而提高模型的识别性能.

表 4 知识数据库对模型识别准确率的影响

Tab.4 Impact of the knowledge databases on model recongnition accuracy

使用情况	在 UCF101 数据集上的识别准确率/%	在 HMDB51 数据集上的识别准确率/%
不使用知识向量数据库	72.5 ± 1.9	51.0 ± 4.9
使用知识向量数据库	73.4 ± 2.1	52.4 ± 4.1

4 结 语

本文针对零样本动作识别任务中先验知识规模不足的问题,提出了一种基于多模态 CLIP 模型与知识数据库的零样本动作识别方法. 为了利用互联网上的非标准化标注数据,以扩充模型的先验知识,使其适应于更广泛的应用场景,本文将图片零样本学习领域的 CLIP 模型迁移到视频任务中,使零样本动作识别模型受益于 CLIP 模型通过自监督对比学习积累的知识. 此外,本文还重新设计了类别语义标签,提高了动作类语义标签中有效信息的含量. 并构建了外部知识数据库,促使模型更有效地建立了视频动作与类别标签之间的联系. 在以上改进的基础上,本文遵循零样本学习的规范对模型进行了微调,以使其更适应于零样本动作识别任务,从而提高了模型的泛化能力和识别性能. 实验结果表明,该方法在 HMDB51 和 UCF101 两个公开数据集上均取得了显著的性能提升,证明了本方法的有效性.

参考文献:

[1] Palatucci M, Pomerleau D, Hinton G E, et al. Zero-shot learning with semantic output codes[C]//Neural Information Processing Systems. Vancouver, Canada, 2009: 1063-6919.

[2] Tian Y, Kong Y, Ruan Q Q, et al. Aligned dynamic-preserving embedding for zero-shot action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(6): 1597-1612.

[3] 冀 中, 郭威辰. 基于局部保持典型相关分析的零样本动作识别[J]. 天津大学学报(自然科学与工程技术版), 2017, 50(9): 975-983.
Ji Zhong, Guo Weichen. Zero shot action reamgnition based on local preserving canonical conelotion analysis[J]. Journal of Tianjin University(Science and Technology), 2017, 50(9): 975-983(in Chinese).

[4] Liu J G, Kuipers B, Savarese S. Recognizing human actions by attributes[C]// The 24th IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, USA, 2011: 3337-3344.

[5] Kodirov E, Xiang T, Fu Z, et al. Unsupervised domain adaptation for zero-shot learning[C]//2015 IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2452-2460.

[6] Roitberg A, Al-Halah Z, Stiefelhagen R. Informed democracy: Voting-based novelty detection for action recognition[EB/OL]. https://arxiv.org/abs/1810.12819, 2018-10-30.

[7] Brattoli B, Tighe J, Zhdanov F, et al. Rethinking zero-shot video classification: End-to-end training for realistic applications[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR). Seattle, USA, 2020: 4612-4622.

[8] Mandal D, Narayan S, Dwivedi S K, et al. Out-of-distribution detection for generalized zero-shot action recognition[C]//2019 IEEE/CVF Conference on Com-

puter Vision and Pattern Recognition(CVPR). Long Beach, USA, 2019: 9977-9985.

[9] Xu X, Hospedales T, Gong S G. Semantic embedding space for zero-shot action recognition[C]//2015 IEEE International Conference on Image Processing(ICIP). Quebec City, Canada, 2015: 63-67.

[10] Gan C, Lin M, Yang Y, et al. Concepts not alone: Exploring pairwise relationships for zero-shot video activity recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 3487-3493.

[11] Zhang B, Hu H X, Sha F. Cross-modal and hierarchical modeling of video and text[C]// Proceedings of the European Conference on Computer Vision(ECCV). Cham, Switzerland, 2018: 374-390.

[12] Zhang C R, Peng Y X. Visual data synthesis via GAN for zero-shot video classification[C]// Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden, 2018: 1128-1134.

[13] Estevam V, Laroca R, Pedrini H, et al. Tell me what you see: A zero-shot action recognition method based on natural language descriptions[J]. Multimedia Tools and Applications, 2024, 83(9): 28147-28173.

[14] Chen S Z, Huang D. Elaborative rehearsal for zero-shot action recognition[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 13618-13627.

[15] 齐秋平. 基于直推式零样本学习的动作识别方法[J]. 计算机科学与应用, 2022, 12(3): 499-507.
Qi Qiuping. Research on transductive zero-shot learning for action recognition[J]. Computer Science and Application, 2022, 12(3): 499-507(in Chinese).

[16] Wang M M, Xing J Z, Liu Y, et al. ActionCLIP: A new paradigm for video action recognition[EB/OL]. https://arxiv.org/abs/2109.08472, 2021-09-17.

[17] Ni B, Peng H W, Chen M H, et al. Expanding language-image pretrained models for general video recognition[C]// European Conference on Computer Vision. Cham, Switzerland, 2022: 1-18.

[18] 吕露露, 黄　毅, 高君宇, 等. 多模态零样本人体动作识别[J]. 中国图象图形学报, 2021, 26(7): 1658-1667.
Lü Lulu, Huang Yi, Gao Junyu, et al. Multimodal-based zero-shot human action recognition[J]. Journal of Image and Graphics, 2021, 26(7): 1658-1667(in Chinese).

[19] Zhu Y, Long Y, Guan Y, et al. Towards universal representation for unseen action recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9436-9445.

[20] Wang Q, Chen K. Alternative semantic representations for zero-shot human action recognition[C]// Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Cham, Switzerland, 2017: 87-102.

[21] Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 7082-7092.

[22] Lin C C, Lin K, Wang L, et al. Cross-modal representation learning for zero-shot action recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 19946-19956.

[23] 兰　红, 方治屿. 零样本图像识别[J]. 电子与信息学报, 2020, 42(5): 1188-1200.
Lan Hong, Fang Zhiyu. Recent advances in zero-shot learning[J]. Journal of Electronics & Information Technology, 2020, 42(5): 1188-1200(in Chinese).

[24] 冯耀功, 于　剑, 桑基韬, 等. 基于知识的零样本视觉识别综述[J]. 软件学报, 2021, 32(2): 370-405.
Feng Yaogong, Yu Jian, Sang Jitao, et al. Survey on knowledge-based zero-shot visual recognition[J]. Journal of Software, 2021, 32(2): 370-405(in Chinese).

[25] 翟永杰, 张智柏, 王亚茹. 基于改进 TransGAN 的零样本图像识别方法[J]. 智能系统学报, 2023, 18(2): 352-359.
Zhai Yongjie, Zhang Zhibai, Wang Yaru. An image recognition method of zero-shot learning based on an improved TransGAN[J]. CAAI Transactions on Intelligent Systems, 2023, 18(2): 352-359(in Chinese).

[26] Pu S, Zhao K L, Zheng M. Alignment-uniformity aware representation learning for zero-shot video classification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 19936-19945.

[27] Gao J Y, Zhang T Z, Xu C S. I Know the relationships: Zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs[C]//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 8303-8311.

[28] Kolesnikov A, Beyer L, Zhai X, et al. Big transfer(BIT): General visual representation learning[C]// European Conference on Computer Vision. Cham, Switzerland, 2020: 491-507.

[29] Wu Z, Fu Y, Jiang Y G, et al. Harnessing object and scene semantics for large-scale video understanding[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Las Vegas, USA, 2016: 3112-3121.

[30] Qi Q P, Wang H L, Su T Y, et al. Learning temporal information and object relation for zero-shot action recognition[J]. Displays, 2022, 73: 102177.

[31] Zhu Y, Li X Y, Liu C H, et al. A comprehensive study of deep video action recognition[EB/OL]. https://arxiv.org/abs/2012.06567, 2020-12-11.

[32] Carreira J, Noland E, Hillier C, et al. A short note on the kinetics-700 human action dataset[EB/OL]. https://arxiv.org/abs/1907.06987v2, 2019-07-15.

[33] Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 4724-4733.

[34] Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild [EB/OL]. https://arxiv.org/abs/1212.0402, 2012-12-03.

[35] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A large video database for human motion recognition[C]// Proceedings of the IEEE International Conference on Computer Vision. Barcelona, Spain, 2011: 2556-2563.

[36] Xu X, Hospedales T, Gong S G. Zero-shot action recognition by word-vector embedding[EB/OL]. https://arxiv.org/abs/1511.04458v1, 2015-11-13.

[37] Lin K, Li L J, Lin C C, et al. SwinBERT: End-to-end transformers with sparse attention for video captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 17928-17937.

[38] Bishay M, Zoumpourlis G, Patras I. TARN: Temporal attentive relation network for few-shot and zero-shot action recognition[EB/OL]. https://arxiv.org/abs/1907.09021v1, 2019-07-21.

[39] Gowda S N. Synthetic sample selection for generalized zero-shot learning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Vancouver, Canada, 2023: 58-67.

(责任编辑:孙立华)