基于语言–视觉对比学习的多模态视频行为识别方法

张 颖[1,2] 张冰冰[1,3] 董 微[1,2] 安峰民[1,2] 张建新[1,2] 张 强[3]

摘 要 以对比语言–图像预训练 (Contrastive language-image pre-training, CLIP) 模型为基础, 提出一种面向视频行为识别的多模态模型, 该模型从视觉编码器的时序建模和行为类别语言描述的提示学习两个方面对 CLIP 模型进行拓展, 可更好地学习多模态视频表达. 具体地, 在视觉编码器中设计虚拟帧交互模块 (Virtual-frame interaction module, VIM), 首先, 由视频采样帧的类别分词做线性变换得到虚拟帧分词; 然后, 对其进行基于时序卷积和虚拟帧分词移位的时序建模操作, 有效建模视频中的时空变化信息; 最后, 在语言分支上设计视觉强化提示模块 (Visual-reinforcement prompt module, VPM), 通过注意力机制融合视觉编码器末端输出的类别分词和视觉分词所带有的视觉信息来获得经过视觉信息强化的语言表达. 在 4 个公开视频数据集上的全监督实验和 2 个视频数据集上的小样本、零样本实验结果, 验证了该多模态模型的有效性和泛化性.

关键词 视频行为识别, 语言–视觉对比学习, 多模态模型, 时序建模, 提示学习

引用格式 张颖, 张冰冰, 董微, 安峰民, 张建新, 张强. 基于语言–视觉对比学习的多模态视频行为识别方法. 自动化学报, 2024, **50**(2): 417−430

DOI 10.16383/j.aas.c230159

Multi-modal Video Action Recognition Method Based on Language-visual Contrastive Learning

ZHANG Ying[1,2] ZHANG Bing-Bing[1,3] DONG Wei[1,2] AN Feng-Min[1,2]
ZHANG Jian-Xin[1,2] ZHANG Qiang[3]

Abstract This paper presents a novel multi-modal model for video action recognition, which is built upon the contrastive language-image pre-training (CLIP) model. The presented model extends the CLIP model in two ways, i.e., incorporating temporal modeling in the visual encoder and leveraging prompt learning for language descriptions of action classes, to better learn multi-modal video representations. Specifically, we design a virtual-frame interaction module (VIM) within the visual encoder that transforms class tokens of sampled video frames into virtual-frame tokens through linear transformation, and then temporal modeling operations based on temporal convolution and virtual-frame token shift are performed to effectively model the spatio-temporal change information in the video. In the language branch, we propose a visual-reinforcement prompt module (VPM) that leverages an attention mechanism to fuse the visual information, carried by the class token and visual token which are both output by the visual encoder, to enhance the language representations. Fully-supervised experiments conducted on four publicly available video datasets, as well as few-shot and zero-shot experiments conducted on two video datasets, demonstrate the effectiveness and generalization capabilities of the proposed multi-modal model.

Key words Video action recognition, language-visual contrastive learning, multi-modal model, temporal modeling, prompt learning

Citation Zhang Ying, Zhang Bing-Bing, Dong Wei, An Feng-Min, Zhang Jian-Xin, Zhang Qiang. Multi-modal video action recognition method based on language-visual contrastive learning. *Acta Automatica Sinica*, 2024, **50**(2): 417−430

收稿日期 2023-03-27 录用日期 2023-08-29
Manuscript received March 27, 2023; accepted August 29, 2023
国家自然科学基金 (61972062), 辽宁省应用基础研究计划 (2023JH2/101300191), 国家民委中青年英才培养计划资助
Supported by National Natural Science Foundation of China (61972062), Applied Basic Research Project of Liaoning Province (2023JH2/101300191), and Young and Middle-aged Talents Program of the National Civil Affairs Commission
本文责任编委 桑农
Recommended by Associate Editor SANG Nong
1. 大连民族大学计算机科学与工程学院 大连 116600 2. 大连民族大学机器智能与生物计算研究所 大连 116600 3. 大连理工大学电子信息与电气工程学部 大连 116024
1. School of Computer Science and Engineering, Dalian Minzu University, Dalian 116600 2. Institute of Machine Intelligence and Bio-computing, Dalian Minzu University, Dalian 116600 3. Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024

视频行为识别是视频理解领域中的重要问题, 其致力于识别视频中人类的不同行为[1−3], 在智能监控、人机交互、医疗健康等众多领域扮演着重要的角色. 进入深度学习时代以来, 以卷积神经网络 (Convolution neural network, CNN)[4−12] 和视觉 Transformer (Vision transformer, ViT)[13−17] 为基

础网络的视频行为识别模型取得了极大发展. 目前, 广泛使用的视频行为识别模型都是在预定义好类别的人工标注数据集上, 以监督学习的方式进行闭集训练. 这类模型只关注视觉表示, 将类别名称转换为矢量标签以简化训练过程, 从而忽略了类别名称的语义信息, 导致学习到的特征对训练数据类别的依赖性高、泛化性差, 识别训练数据类别以外的视频需要重新标注数据以提供监督信号并进行额外的训练. 为实现学习视觉通用表示来解决各种现实问题的目标, 这种面向固定类别、只能解决单一问题的模型显然不能满足需求.

最近, 语言−视觉对比学习模型[18−20] 开拓了解决闭集训练问题的思路, 给学习泛化性能更强的通用视觉表示带来了希望, 尤其是在零样本学习上表现出较大潜力. 这类模型保留类别标签的语言描述作为监督信号, 将视觉单模态模型拓展到语言−视觉多模态架构, 在亿级甚至 10 亿级别的语言−图像对上进行自监督训练, 以对比学习的方式同时优化语言和视觉编码器. 在测试过程中, 该对比学习模型直接将单模态架构下的概率模型转换为语言−图像的检索问题. 受到语言−视觉对比学习模型的启发, 本文尝试在视频行为识别任务中引入语言监督来辅助学习更广泛的视觉概念, 以获得泛化性更强的视频表达.

首先, 考虑到视频训练所需计算成本较高, 以及现有视频数据集规模相对较小的问题, 抛弃在语言−视频数据上从头进行预训练的做法, 致力于解决如何将对比语言−图像预训练 (Contrastive language-image pre-training, CLIP)[18] 模型迁移到视频行为识别任务的问题. 其次, 由于视频存在区别于图像的时序信息, 并且其类别标签所携带的语言信息十分有限, 完成 CLIP 模型从图像到视频的迁移和适应必须解决两个关键问题: 1) 如何利用视频的时序信息; 2) 如何强化现有类别标签的语言表达. 为此, 提出一种新的时序建模结构, 将图像编码器拓展为视频编码器. 该结构包含虚拟帧交互模块和全局帧融合模块两个关键部件. 虚拟帧交互模块借助“虚拟帧分词”完成帧间信息交互, 使帧级编码器能提供包含时序信息的帧级表达; 全局帧融合模块集成帧级特征, 得到视频表达. 同时, 针对问题 2) 提出视觉强化提示模块, 该模块通过注意力机制融合视觉信息, 自动生成语言提示, 将初级语言表达转换为强化语言表达. 依赖上述两方面拓展, CLIP 模型可以有效适用于视频行为识别任务. 本文主要贡献如下:

1) 在 CLIP 模型基础上, 提出一种新颖的基于语言−视觉对比学习的多模态视频行为识别模型, 将 CLIP 模型在图像领域的先验知识迁移至视频领域.

2) 对于视觉编码器, 提出虚拟帧交互模块. 该模块首先使用视频采样帧的类别分词进行线性变换生成“虚拟帧分词”, 接着通过对“虚拟帧分词”进行时序卷积和虚拟帧分词移位操作完成帧间信息交互, 实现在网络中间层充分利用视频时序信息的目的.

3) 对于语言分支, 提出视觉强化提示模块. 该模块通过注意力机制融合视频编码器输出的类别分词和视觉分词中带有的视觉信息, 自动生成适应视频行为识别的语言提示, 对语言编码器生成的初级语言表达进行加权来达到强化语言表达的目的.

4) 在 4 个视频行为识别公开数据集上进行全监督、小样本和零样本实验, 验证本文模型的有效性和泛化性.

1 相关工作

1.1 CLIP 在视频行为识别上的应用

CLIP 是语言−视觉对比学习的代表性工作之一, 其核心思想是直接从图片标签对应的语言描述中获取额外的监督信号, 通过对比学习产生泛化性能强大的视觉表达. 文献 [21−27] 将 CLIP 应用于视频行为识别, 表现出十分有前景的性能. 这些工作可大致分为两类: 1) 专注视觉建模的单模态模型仅使用 CLIP 图像编码器的预训练参数对视频编码器做强初始化, 如 ST-Adaptor (Spatio-temporal adapter)[21]、EVL (Efficient video learning)[22] 和 AIM (Adapt pre-trained image models)[23]; 2) 继承 CLIP 的多模态思想和“双塔”架构的语言−视频多模态模型, 如 VideoCLIP[24]、PromptCLIP[25]、ActionCLIP[26] 和 X-CLIP[27].

文献 [21−23] 在单模态框架下利用 CLIP 预训练模型, 虽然这类方法可以通过参数初始化的方式利用多模态预训练的优势, 特别是 AIM 在纯图像主干下简单地重用预训练的自注意进行时间建模, 取得了优秀的性能, 但是使模型失去了在微调过程中进一步进行多模态交互的能力. 为更好地探索多模态交互对模型表达能力的促进作用, 本文选择在多模态框架下进行模型迁移. 在多模态框架下, VideoCLIP 将语言−图像数据替换为语言−视频数据, 从头进行预训练. 这种方式对存储硬件、计算资源的要求很高, 而且实验周期相对更长. PromptCLIP 冻结预训练参数, 仅通过优化文本分支的一组可学习的“连续提示向量”和位于视觉编码器末端的轻量级时序 Transformer, 来使预先训练的语言−视觉模

型适应视频理解任务. 其提示学习的信息仅在语言分支流动, 且对时序信息的利用不够充分. ActionCLIP 提出一个“预训练–提示–微调”的范式, 通过手工设计的语言提示和视觉提示, 将 CLIP 模型的知识迁移到视频行为识别. X-CLIP 提出一种新的跨帧通信注意力操作和特定于视频的提示学习方法, 来完成 CLIP 模型从图像到视频的扩展. 本文模型与 ActionCLIP 和 X-CLIP 最为相关, 首先, 将 CLIP 模型从二维图像域扩展到三维视频域; 然后, 在视频行为识别基准上进行端到端的微调, 来达到知识迁移的目的.

相比之下, 本文模型是性能与计算成本权衡下的更优选择. 其中, 虚拟帧交互模块作为一种“即插即用”的轻量级时序建模模块, 克服了在网络中间层融入时序建模可能导致的“灾难性遗忘”风险. 相较于 ActionCLIP 在视觉分支末端使用 Transformer 层融合时序信息的做法, 本文通过在模型中间层插入虚拟帧交互模块, 更充分地利用了视频时序信息. 视觉强化提示模块同时将视频编码器输出的类别和视觉分词作为提示信号, 利用其中包含的视觉信息自动生成语言提示, 得到强化语言表达. 相较于 X-CLIP 中的特定于视频的提示学习方法, 视觉强化提示模块更充分地考虑了不同视频分词包含的视觉信息对语言表达强化作用的互补性.

1.2 时序建模方式

视频由一系列静态图像组成, 这些视频帧不是简单的图像集合, 它们之间存在前后关联的时序关系. 因此, 精准、有效地对视频帧之间的时序信息进行建模是建立具有鲁棒性视频行为识别模型的关键因素. 进入深度学习时代以来, CNN 在计算机视觉领域占有主导地位. 基于 CNN 的网络架构下有 3 类代表性的时序建模方式: 1) 添加额外的分支对运动信息 (光流) 建模, 以描述帧间的时序关系[4]; 2) 使用 3D 卷积[5−6] 将时间维度纳入计算或将 3D CNN 在时间和空间维度上进行分解, 以提高计算效率[7−9]; 3) 在 2D CNN 的基础网络上, 插入时序建模模块[10−12] 来平衡识别准确率和计算速度.

基于 CNN 的网络虽然在视频行为识别任务上取得了优秀的成绩, 但卷积计算的有限感受野限制了模型对远距离依赖的建模能力. Transformer[28] 应用注意力机制将序列中任意两个位置之间的距离缩小为一个常量, 有效解决了远距离依赖建模的难题. ViT[29] 将 Transformer 从自然语言处理迁移至计算机视觉后, 依赖性能优秀的预训练模型, 基于 ViT 的模型开始在视频领域被广泛使用. 由于注意力计算的高复杂度, 在 Transformer 架构下, 直接进行时空联合注意力计算显然会带来过高的计算成本. 因此, 存在两类平衡识别准确率和计算复杂度的方式: 1) 设计计算更节约的网络架构. 如 VTN (Video transformer network)[13] 和 ViViT (Video vision transformer)[14] 划分空间和时间编码, 先用帧级编码器对视频帧进行独立的空间建模, 再利用一个 Transformer 层融合时间信息. 2) 设计计算开销更小的新的注意力机制. 如文献 [15] 将时空联合注意力机制分为空间和时间的注意力计算, MViT (Multi-scale vision transformers)[16] 提出多尺度视觉 Transformer, Swin (Video swin transformer)[17] 在视频 Transformer 中引入局部归纳偏置. 本文的时序建模参考 VTN 和 ViViT, 先进行帧级编码, 再进行时间融合. 与在各帧上进行独立空间编码不同的是, 本文虚拟帧交互模块在允许帧级编码器在空间编码的同时, 充分利用视频中的时序线索.

1.3 提示学习

提示学习源于自然语言处理, 其通过给予适当的“提示”, 让模型更好地理解下游任务的需求, 并产生相应的输出. 例如给定一个示例后, GPT-3 (Generative pre-trained transformer 3)[30] 模型即使某个目标领域从未被训练过, 也可以自动给出该目标领域的答案. 根据构建方式, 可将提示学习分为两类: 1) 手工模板工程设计精准, 但对领域知识依赖性高且灵活性差; 2) 自动模板学习包括离散提示[31−34] 和连续提示[35−36]. 对于手工模板工程, “提示”是实际的语言字符串; 在自动模板学习中, “提示”被映射到一个连续的向量空间中, 不是可解释的自然语言. 计算机视觉同样受益于提示学习, CLIP 和 ActionCLIP 在测试阶段使用手工模板辅助模型表现出强大的泛化能力. CoOp (Context optimization)[37] 和 CoCoOp (Conditional context optimization)[38] 开始探索可学习的语言提示方法. 本文延续连续提示方法的思想, 提出一种可学习的自动语言提示生成方法, 称为“视觉强化提示”. 该方法将视频编码器输出的类别分词和视觉分词作为提示信号, 通过注意力机制融合其带有的视觉信息, 自动生成适应视频行为识别的语言提示, 以强化语言表达.

2 基于语言–视觉对比学习的多模态模型

本文从时序建模和提示学习两个方面对 CLIP 模型进行拓展, 得到一种基于语言–视觉对比学习的多模态模型, 该模型有效适用于视频行为识别任务. 本节安排如下: 第 2.1 节概述模型结构, 第 2.2

节和第 2.3 节介绍本文提出的虚拟帧交互模块和视觉强化提示模块.

2.1 模型结构概述

如图 1 所示, 模型遵循视觉分支和语言分支独立的"双塔"结构, 通过联合训练视频编码器和语言编码器学习对齐视频表达和相应的语言表达. 具体地, 模型整体包含视频编码器 $\mathcal{F}_{\theta_v}(\cdot)$、语言编码器 $\mathcal{F}_{\theta_t}(\cdot)$ 和视觉强化提示模块 $\mathcal{F}_{\theta_p}(\cdot)$ 三个组成部分. 给定一个语言–视频数据集 $(\mathcal{L}, \mathcal{V})$, $V \in \mathcal{V}$ 是一个视频片段, $L \in \mathcal{L}$ 是其对应类别标签的语言描述. 将 V 和 L 分别输入视频编码器 $\mathcal{F}_{\theta_v}(\cdot)$ 和语言编码器 $\mathcal{F}_{\theta_t}(\cdot)$ 中, 获得视频表达 $\boldsymbol{v}$ 和初级语言表达 $\boldsymbol{\ell}$:

$$\boldsymbol{v}, \boldsymbol{v}_{cls}, \boldsymbol{v}_{vis} = \mathcal{F}_{\theta_v}(V) \tag{1}$$

$$\boldsymbol{\ell} = \mathcal{F}_{\theta_t}(L) \tag{2}$$

式中, $\boldsymbol{v}_{cls}$ 和 $\boldsymbol{v}_{vis}$ 分别为视频编码器输出的类别分词和视觉分词. 将两类视频分词作为视觉提示信号, 与初级语言表达 $\boldsymbol{\ell}$ 共同输入视觉强化提示模块 $\mathcal{F}_{\theta_p}(\cdot)$, 获得经过视觉信息强化的语言表达 $\hat{\boldsymbol{\ell}}$:

$$\hat{\boldsymbol{\ell}} = \mathcal{F}_{\theta_p}(\boldsymbol{v}_{cls}, \boldsymbol{v}_{vis}, \boldsymbol{\ell}) \tag{3}$$

最后, 模型会计算视频表达 $\boldsymbol{v}$ 与强化语言表达 $\hat{\boldsymbol{\ell}}$ 之间的相似性:

$$\cos\left(\boldsymbol{v}, \hat{\boldsymbol{\ell}}\right) = \frac{\langle \boldsymbol{v}, \hat{\boldsymbol{\ell}} \rangle}{||\boldsymbol{v}||||\hat{\boldsymbol{\ell}}||} \tag{4}$$

式中, $\cos(\cdot)$ 代表余弦相似度计算. 数据集中一一对应的语言–视频对被视作正样本, 其他组合均视作负样本. 模型最终的优化目标是当 V 和 L 互为正样本时, 最大化相似度计算值; 反之, 最小化相似度计算值.

2.1.1 视觉分支

模型的视觉分支由视频编码器 $\mathcal{F}_{\theta_v}(\cdot)$ 构成, 视频编码器包含帧级编码器 $\mathcal{H}(\cdot)$ 和全局帧融合模块两个组成部分. 首先, 帧级编码器以原始帧作为输入, 提供帧级表达; 然后, 全局帧融合模块集成帧级表达, 得到视频表达.

具体地, 给定一个视频片段 $V \in \mathbf{R}^{T\times H\times W\times 3}$, 其中 T 为采样帧数, $H \times W$ 为视频帧的空间分辨率, 输入通道数为 3. 首先, 依据 ViT 对图像的划分方法将各视频帧独立划分为 N 个互不重叠的图像块 $\{\boldsymbol{x}_{t,i}\}_{i=1}^{N} \in \mathbf{R}^{P^2\times 3}$, 空间分辨率为 $P \times P$. 其中 $t \in \{1, 2, \cdots, T\}$ 代表视频帧的时序索引, $N = HW/P^2$ 代表第 t 个视频帧被划分的图像块数量. 然后, 使用线性映射 $\boldsymbol{E} \in \mathbf{R}^{3P^2\times D}$ 将各帧图像块 $\{\boldsymbol{x}_{t,i}\}_{i=1}^{N} \in \mathbf{R}^{P^2\times 3}$ 映射为序列长度为 N 的块嵌入, 并在各帧序列首部拼接一个可学习的类别分词 $(\boldsymbol{z}_{t,0} = \boldsymbol{x}_{class})$.

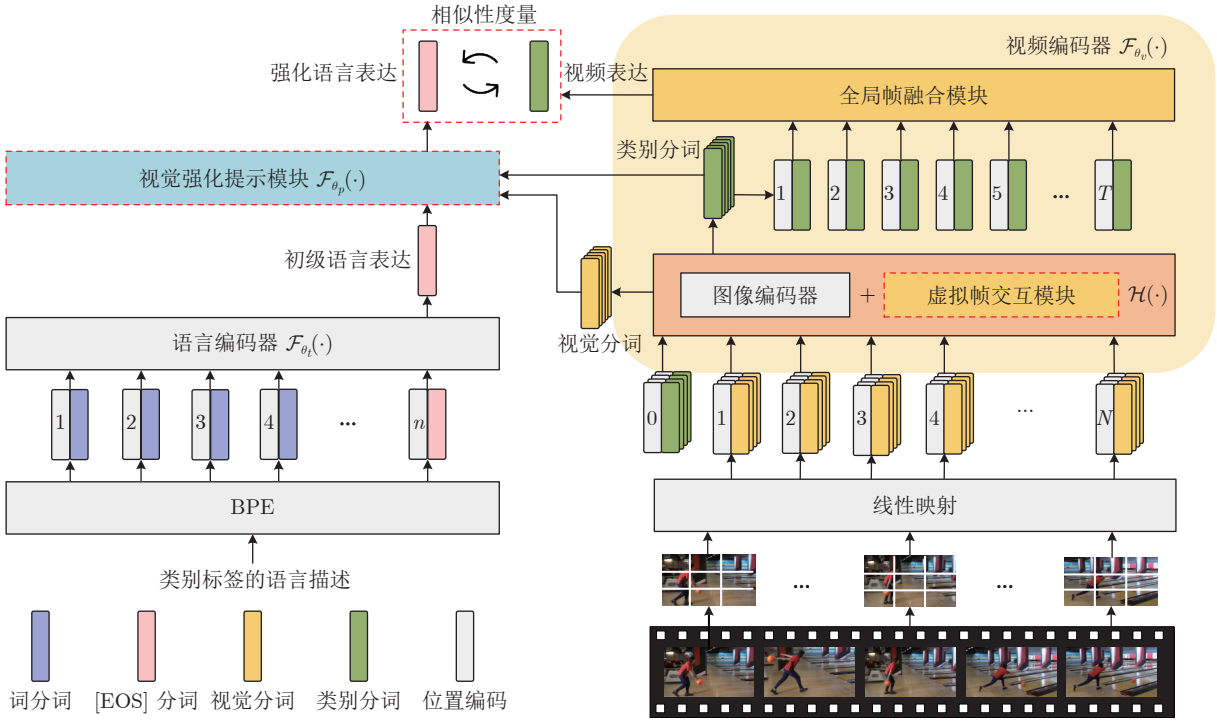

图 1 基于语言–视觉对比学习的多模态模型

Fig. 1 Multi-modal model based on language-visual contrastive learning

最后, 给各帧序列添加空间位置编码 $\boldsymbol{e}_{spa} \in \mathbf{R}^{(N+1)\times D}$ 以保留图像块在原始帧中的位置信息. 帧级编码器 $\mathcal{H}(\cdot)$ 在第 t 帧处的输入可表示为:

$$\boldsymbol{z}_t^{(0)} = [\boldsymbol{x}_{class}, \boldsymbol{E}\boldsymbol{x}_{t,1}, \boldsymbol{E}\boldsymbol{x}_{t,2}, \cdots, \boldsymbol{E}\boldsymbol{x}_{t,N}] + \boldsymbol{e}_{spa} \tag{5}$$

将式 (5) 中的 $\boldsymbol{z}_t^{(0)} \in \mathbf{R}^{(N+1)\times D}$ 输入帧级编码器 $\mathcal{H}(\cdot)$, 得到帧级表达 $\boldsymbol{h}_t$:

$$\boldsymbol{z}_t^{(l)} = \mathcal{H}\left(\boldsymbol{z}_t^{(l-1)}\right), \quad l = 1, \cdots, L_{\mathcal{H}} \tag{6}$$

$$\boldsymbol{h}_t = \boldsymbol{z}_{t,0}^{(L_{\mathcal{H}})} \tag{7}$$

式中, l 代表 $\mathcal{H}(\cdot)$ 的层索引编号, 共有 $L_{\mathcal{H}}$ 层. $\boldsymbol{z}_{t,0}^{(L_{\mathcal{H}})}$ 是类别分词在输出序列中对应的输出状态. 构成帧级编码器的虚拟帧交互模块在第 2.2 节详细描述.

最后, 将整个视频的帧级表达的集合 $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_T]$ 输入全局帧融合模块 (Global frame fusion module, GBM), 进行时序上的"后融合", 以生成视频表达 $\boldsymbol{v}$:

$$\boldsymbol{v} = \text{AvgPool}\left(\text{GBM}\left(\boldsymbol{H} + \boldsymbol{E}_{temp}\right)\right) \tag{8}$$

式中, AvgPool 和 $\boldsymbol{E}_{temp} \in \mathbf{R}^{T\times D}$ 分别代表平均池化和时间位置编码. GBM 由标准的多头自注意力 (Multi-head self-attention, MHSA) 和前馈神经网络 (Feed-forward networks, FFN) 组成.

在视觉分支中, 对虚拟帧交互模块和全局帧融合模块做随机初始化, 其他部分使用 CLIP 预训练模型权重.

2.1.2　语言分支

模型的语言分支由语言编码器 $\mathcal{F}_{\theta_t}(\cdot)$ 和视觉强化提示模块构成. 首先, 语言编码器将类别标签的语言描述作为输入, 提供初级语言表达; 然后, 视觉强化提示模块将初级语言表达转换为强化语言表达.

本文使用预训练的 CLIP 语言编码器 (含 12 层 Transformer) 生成视频的初级语言表达. 具体地, 首先, 对类别标签的语言描述进行字节对编码 (Byte pair encoding, BPE)[39], 得到语言序列 $\{\boldsymbol{s}_i\}_{i=1}^{n}$. 为方便批量处理, 将语言序列的长度 n 固定为 77, 包含在序列尾部添加的 [EOS] 分词. 然后, 对语言序列进行词嵌入, $d = 512$ 代表语言序列的嵌入维度. 语言编码器 $\mathcal{F}_{\theta_t}(\cdot)$ 的输入可表示为:

$$\boldsymbol{c}^{(0)} = \text{Embed}\left[\boldsymbol{s}_1, \boldsymbol{s}_2, \cdots, \boldsymbol{s}_n\right] + \boldsymbol{e} \tag{9}$$

式中, Embed 代表词嵌入操作, $\boldsymbol{e} \in \mathbf{R}^{n\times d}$ 是分词的位置编码. $\boldsymbol{c}^{(0)} \in \mathbf{R}^{n\times d}$ 被输入到语言编码器 $\mathcal{F}_{\theta_t}(\cdot)$ 中, 得到初级语言表达 $\boldsymbol{\ell}$:

$$\boldsymbol{c}^{(l)} = \mathcal{F}_{\theta_t}\left(\boldsymbol{c}^{(l-1)}\right), \quad l = 1, \cdots, 12 \tag{10}$$

$$\boldsymbol{\ell} = \boldsymbol{c}_{index([\text{EOS}])}^{(12)} \tag{11}$$

式中, l 代表 $\mathcal{F}_{\theta_t}(\cdot)$ 的层索引编号, $index([\text{EOS}])$ 代表 [EOS] 分词的索引编号. $\boldsymbol{c}_{index([\text{EOS}])}^{(12)}$ 是 [EOS] 分词在输出序列中对应的输出状态.

最后, 如式 (3) 所示, 初级语言表达 $\boldsymbol{\ell}$ 和两类视频分词共同输入视觉强化提示模块, 得到最终用于视频行为识别的强化语言表达 $\hat{\boldsymbol{\ell}}$. 视觉强化提示模块在第 2.3 节详细描述. 在语言分支中, 视觉强化提示模块进行随机初始化, 语言编码器使用预训练模型权重.

2.2　虚拟帧交互模块

考虑到仅在帧级编码器末端以"后融合"方式进行时序建模的做法[13–14] 未能充分利用视频的时序信息, 本文提出虚拟帧交互模块 (Virtual-frame interaction module, VIM). 该模块使帧级编码器能够在网络中间层进行空间编码的同时, 融合视频的时序信息.

在网络中间层插入时序建模模块, 可能会破坏预训练模型的表达能力. 例如 ActionCLIP 在网络中间层引入时序移动模块[10], 引发了"灾难性遗忘"问题. 本文认为其原因在于对特征做全局移动操作导致了相对预训练模型的大幅数据分布偏差. 因此, 如图 2 所示, 虚拟帧交互模块为每一帧引入一个"虚拟帧分词", 来完成帧间信息交换. 具体地, 首先, 第 l 层中第 t 帧的虚拟帧分词 $\boldsymbol{f}_t^{(l)}$ 由第 $l-1$ 层的类别分词 $\boldsymbol{z}_{t,0}^{(l-1)}$ 进行线性变换得到, 在形式上代表该帧在当前层的抽象信息, 该模块对全部虚拟帧分词进行基于时序卷积 (Temporal convolution, T-Conv) 和虚拟帧分词移位 (Virtual-frame token shift, VT-Shift) 的时序建模操作. T-Conv 是应用于时间维度的一维卷积, 使用大尺寸卷积核以捕获视频远距离依赖. 受到 Token Shift[40] 沿时间维度移动部分类别分词进行相邻帧信息交互的做法的启发, VT-Shift 沿时间维度移动部分虚拟帧分词以达到相邻帧信息交换的目的. 这样, 连续的 T-Conv 和 VT-Shift 所构成的 VIM 就兼顾了相邻帧和远距离的时序依赖建模. 在计算成本上, VT-Shift 是一个 0 参数量、0 计算量的高效操作; T-Conv 在实现时使用分组卷积, 以平衡使用较大卷积核所带来的计算开销. 其中, 第 l 层的虚拟帧交互模块的计算过程可表示为:

$$\hat{\boldsymbol{F}}^{(l)} = \boldsymbol{F}^{(l)} + \text{VT-Shift}\left(\delta\left(\text{LN}\left(\text{T-Conv}\left(\boldsymbol{F}^{(l)}\right)\right)\right)\right) \tag{12}$$

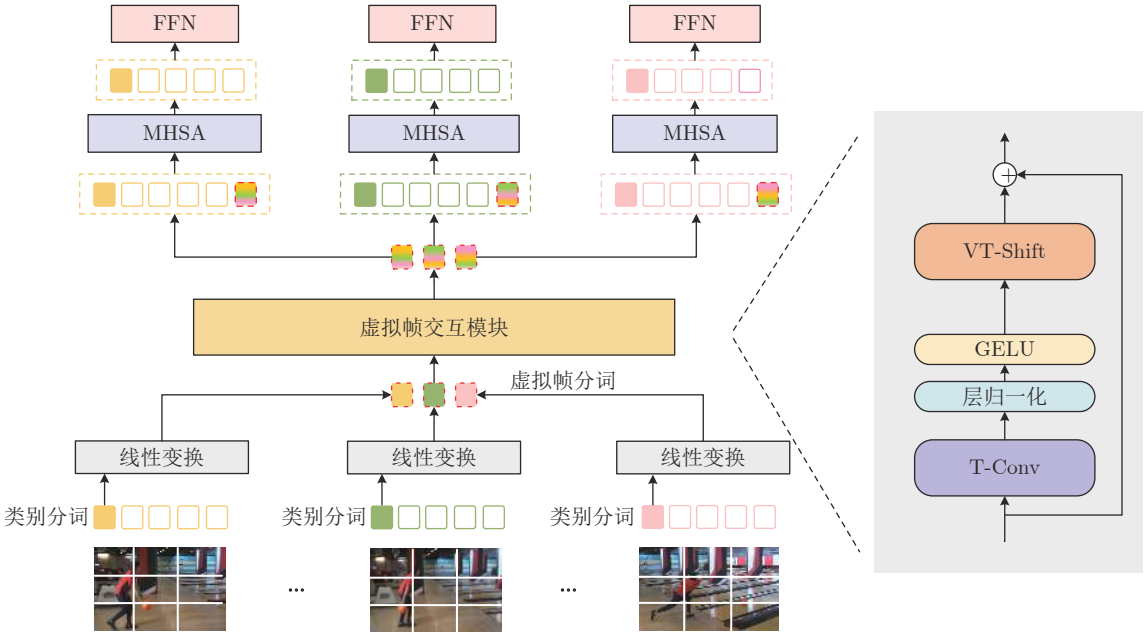

图 2 虚拟帧交互模块

Fig. 2 Virtual-frame interaction module

式中, $\hat{\boldsymbol{F}}^{(l)} = [\hat{\boldsymbol{f}}_1^{(l)}, \hat{\boldsymbol{f}}_2^{(l)}, \cdots, \hat{\boldsymbol{f}}_T^{(l)}]$, LN 代表层归一化操作, δ 代表激活函数 GELU.

然后, 虚拟帧交互模块输出的携带时序信息的虚拟帧分词与对应帧的视觉分词拼接, 共同输入标准的 MHSA 和 FFN, 进行帧内空间建模:

$$\left[\hat{\boldsymbol{z}}_t^{(l)}, \tilde{\boldsymbol{f}}_t^{(l)}\right] = \left[\boldsymbol{z}_t^{(l-1)}, \hat{\boldsymbol{f}}_t^{(l)}\right] + \mathrm{MHSA}\left(\mathrm{LN}\left(\left[\boldsymbol{z}_t^{(l-1)}, \hat{\boldsymbol{f}}_t^{(l)}\right]\right)\right) \tag{13}$$

$$\boldsymbol{z}_t^{(l)} = \hat{\boldsymbol{z}}_t^{(l)} + \mathrm{FFN}\left(\mathrm{LN}\left(\hat{\boldsymbol{z}}_t^{(l)}\right)\right) \tag{14}$$

式中, $[\cdot, \cdot]$代表虚拟帧分词与视觉分词拼接, 虚拟帧分词在进入 FFN 层之前被丢弃. 值得注意的是, 虚拟帧分词携带的时间信息在 MHSA 和 FFN 中得以继续扩散和加强, 这部分计算使用空间建模的计算成本建模了全局时空依赖.

虚拟帧交互模块保持原始帧嵌入不变, 借助“虚拟帧分词”进行时序建模的做法避免了对帧级特征做全局移动可能导致的“灾难性遗忘”风险. 作为一个轻量级时序建模模块与其他时序建模机制[14−16, 27]相比, 大大降低了计算成本. 同时, 虚拟帧交互模块可插入任何基于 Transformer 的视觉编码器中, 在网络的中间层捕获帧间时序信息及远距离依赖, 是一种“即插即用”的中间层时序建模模块.

2.3 视觉强化提示模块

现有视频数据集的类别标签大多只是一个单词或短语, 不充分的语言信息限制了预训练模型的表达能力. 为强化语言表达, 本文提出视觉强化提示模块 (Visual-reinforcement prompt module, VPM), 该模块借助视觉信息, 自动生成适应视频行为识别的语言, 提示以强化语言表达.

参考人类理解图像时的语言、视觉多模态协作, 语言−视觉多模态模型[18−20]取得了巨大成功. 但是, 目前这种多模态协作大多体现在语言对视觉的监督作用, 而在语言分支采用离线的手工提示模板[18, 26]或给语言分词拼接可学习随机向量[37]的做法忽略了视觉信息对语言表达的强化作用. 基于这点思考, 本文设计了视觉强化提示模块. 该模块通过注意力机制融合视频编码器输出的视频分词中带有的视觉信息, 自动生成适应视频行为识别的语言提示, 从视觉对语言的强化作用的角度, 深化了语言与视觉的多模态交互. 在视频分词的选择上, 使用类别分词对应的输出作为视觉信息的代表进行分类, 是一种被广泛认可和使用的做法. SoT (Second-order transformer)[41] 认为视觉分词中包含的丰富图像语义是对类别分词的补充, 并同时使用视觉分词和类别分词进行分类. 受到 SoT 的启发, VPM 同时使用视觉分词和类别分词作为提示信号, 与语言信息进行跨模态交互来更充分地利用不同视频分词中包含的视觉信息.

具体地, 视觉强化提示模块的每个块都由多头交叉注意力 (Multi-head cross attention, MHCA) 和 FFN 组成, 以初级语言表达 $\boldsymbol{\ell}$ 作为查询信号, 代表视频帧级表达的类别分词 $\boldsymbol{H} \in \mathbf{R}^{T\times D}$ 和代表视频

各帧内容的视觉分词 $\bar{\boldsymbol{z}}$ 分别作为两组键和值, 给予语言表达直接的视觉提示信号, 生成两类适应视频行为识别的语言提示 $\tilde{\boldsymbol{\ell}}_{cls}$ 和 $\tilde{\boldsymbol{\ell}}_{vis}$:

$$\begin{cases}\bar{\boldsymbol{\ell}}_{cls}=\boldsymbol{\ell}+\mathrm{MHCA}\left(\boldsymbol{\ell},\ \boldsymbol{H}\right)\\ \tilde{\boldsymbol{\ell}}_{cls}=\bar{\boldsymbol{\ell}}_{cls}+\mathrm{FFN}\left(\bar{\boldsymbol{\ell}}_{cls}\right)\end{cases}\tag{15}$$

$$\begin{cases}\bar{\boldsymbol{\ell}}_{vis}=\boldsymbol{\ell}+\mathrm{MHCA}\left(\boldsymbol{\ell},\bar{\boldsymbol{z}}\right)\\ \tilde{\boldsymbol{\ell}}_{vis}=\bar{\boldsymbol{\ell}}_{vis}+\mathrm{FFN}\left(\bar{\boldsymbol{\ell}}_{vis}\right)\end{cases}\tag{16}$$

式中, $\bar{\boldsymbol{z}}\in\mathbf{R}^{N\times D}$ 由 $\{\boldsymbol{z}_t^{(L_{\mathcal{H}})}\}_{t=1}^{T}$ 在时间维度上求平均得到. 最后, 将两类语言提示与初级语言表达 $\boldsymbol{\ell}$ 进行加权, 达到强化语言表达的目的:

$$\hat{\boldsymbol{\ell}}=\boldsymbol{\ell}+\alpha\tilde{\boldsymbol{\ell}}_{cls}+\beta\tilde{\boldsymbol{\ell}}_{vis}\tag{17}$$

式中, α 和 β 是两个可学习参数, $\hat{\boldsymbol{\ell}}$ 是最终用于与视频表达计算相似性的强化语言表达. 视觉强化提示模块分别使用视频类别和视觉分词构建原始语言描述以外的视频上/下文语言提示, 充分地考虑了不同视频分词所包含的视觉信息对语言表达强化作用的互补性. 同时, 相较于手工设计模板和可学习随机向量需要在前期的语言符号化过程中完成复杂设计的特点, 该模块直接置于语言编码器末端, 将提示学习融入了模型整体设计.

3　实验

3.1　实验设置

3.1.1　模型实例

本文模型的语言编码器与 CLIP 一样采用 12 层 Transformer ($L_c=12$, $N_h=8$, $d=512$). 其中, L_c 表示层数, N_h 表示注意力头数. 模型采用 ViT (B) ($L_c=12$, $N_h=12$, $d=768$) 作为基础网络构建帧级编码器, 设置 32 × 32 像素和 16 × 16 像素两种视频帧分块尺寸. 全局帧融合模块 ($L_c=1$, $N_h=8$, $d=512$) 和视觉强化提示模块 ($L_c=2$, $N_h=8$, $d=512$), 层数分别设置为 1 和 2.

3.1.2　数据集

本文使用 K400 (Kinetics-400)[42]、Mini-K200 (Mini Kinetics-200)[9]、UCF101[43] 和 HMDB51[44] 四个公开视频行为识别数据集, 对本文提出模型的有效性和泛化性进行验证和对比. 其中, K400 是一个用于视频行为识别的大规模、高质量数据集, 共有 400 个动作类别, 如人与物体的交互、人与人的交互, 包含约 240 000 个训练视频和约 20 000 个测试视频, 每个视频片段大约持续 10 s; Mini-K200 是 K400 的子集, 包含 200 个动作类别, 每个类别分别从训练集和测试集中随机抽取 400 个和 25 个视频, 共产生 80 000 个训练视频和 5 000 个测试视频; UCF101 包含约 13 000 个视频片段, 视频固定帧率为 25 fps, 分为 101 个动作类别, 涵盖人类身体运动、人与人的互动、人与物的互动、演奏乐器和体育运动 5 个组别; HMDB51 包含来自 51 个动作类别的约 6 000 个视频片段, 每个类别至少包含 101 个视频. UCF101 和 HMDB51 分别提供 3 种不同的训练和测试数据的划分方式, 用于衡量模型性能.

3.1.3　实验细节

在训练过程中, 采用与 TSN (Temporal segment network)[45] 相同的视频采样方法, 将视频片段分为 T 个部分, 并在每个部分中随机采样 1 帧, 设置采样帧数 $T=8$ 或 $T=16$. 采样帧输入尺寸为 224×224 像素. 与 X-CLIP 相同, 在训练过程中, 采用正则化和多种数据增强策略, 详细的模型训练超参数见附录 A. 在测试过程中, 全监督实验采用与 ViViT 和 Swin 相同的 4 个时间剪辑 × 3 个空间裁剪的多视图测试, 即对一个视频在时间维度上进行 4 次帧采样, 每个采样帧取 3 个不同的空间裁剪, 取 12 个视图识别结果的平均值作为最终结果. 小样本和零样本实验采用 1 个时间剪辑 × 1 个空间裁剪的单视图测试, 仅对视频做 1 次帧采样, 对采样帧做中心裁剪. 3 种实验设置均以预测概率排名第 1 的类别与实际结果相符的准确率 (称为第 1 识别准确率) 或预测概率排名前 5 的类别包含实际结果的准确率 (称为前 5 识别准确率) 作为评价标准. 对于 HMDB51 和 UCF101 数据集, 在全监督和零样本实验中, 采用 3 种划分方式的测试集第 1 识别准确率的平均值作为评价指标; 在小样本实验中, 仅使用第 1 种划分方式进行训练和测试. 无特殊说明, 在所有消融实验中, 模型均使用分块尺寸为 32 × 32 像素、8 帧输入、单视图测试.

本文模型的训练和测试基于开源的 PyTorch 深度学习框架, 该框架在配备 2 个 NVIDIA GeForce RTX 3090 GPU 的服务器上运行.

3.2　消融实验

在 Mini-K200 上, 对本文模型的各个模块和设置进行充分的消融实验和分析.

3.2.1　语言信息的作用

为评估引入语言信息的作用, 设计一种单模态模型的变体, 该变体与本文模型具有相同的视觉基础网络、时序建模方法和预训练权重, 将语言编码器替换为随机初始化的全连接层作为分类头. 如表 1 所示, 本文模型比对应的单模态模型变体的识别准确率提升 3.0% 和 3.2%. 这表明, 语言中所包含的语义信息有助于模型学习更具判别性的视频表

表 1 模型具有/不具有语言信息的消融研究 (%)
Table 1 Ablation studies of the model w/wo language information (%)

方法	第 1 识别准确率	前 5 识别准确率
单模态变体	82.7	94.0
本文模型	85.7 **(提升 3.0)**	97.2 **(提升 3.2)**

达, 可以有效提升模型性能.

3.2.2 各模块的消融实验

为了验证本文提出模型和各模块的有效性, 设计一个简单的基线 CLIP-Mean. 该基线使用语言监督, 但不包含任何时序建模和提示学习方法, 将帧级表达做简单平均获得视频表达. 依次在 CLIP-Mean 上添加各模块, 观察性能提升情况, 如表 2 所示. 首先, 在基线的图像编码器中插入虚拟帧交互模块 VIM, 第 1 识别准确率提升 0.6%; 接着, 在帧级编码器末端添加全局帧融合模块 GBM, 识别准确率进一步提升 0.2%, 说明本文提出的模型能通过虚拟帧交互模块和全局帧融合模块有效利用视频的时序线索, 且二者的时序建模作用互补; 最后, 继续在语言编码器末端添加视觉强化提示模块 VPM, 识别准确率进一步提升 0.9%, 说明视觉强化提示模块借助视觉信息生成的语言提示有效强化了原有的视频语言表达, 使之更具判别性. 总之, 本文模型相较基线模型, 提升了 1.7%, 成功适应视频行为识别, 验证了它的有效性.

表 2 本文模块的消融实验结果 (%)
Table 2 Ablation studies of the proposed modules (%)

模块	第 1 识别准确率
基线 (CLIP-Mean)	84.0
基线 + VIM	84.6 **(提升 0.6)**
基线 + VIM + GBM	84.8 **(提升 0.8)**
基线 + VIM + GBM + VPM	85.7 **(提升 1.7)**

3.2.3 虚拟帧交互模块超参数的选择

虚拟帧交互模块由时序卷积 T-Conv 和虚拟帧分词移位 VT-Shift 两个时序建模操作构成. 为评估 T-Conv 的卷积核尺寸和 VT-Shift 的移动比例对时序建模效果的影响, 仅在基线上添加虚拟帧交互模块, 依次进行 VT-Shift (移动比例分别为 {1/4, 1/8, 1/16}) 和 T-Conv (卷积核尺寸分别为 {3, 5, 7} 像素) 操作. 图 3 展示了对 VT-Shift 和 T-Conv 的超参数进行消融实验的结果.

如图 3(a) 所示, 对虚拟帧分词施加适当移动比例的 VT-Shift 操作, 可捕获相邻帧之间的时序信息, 当移动比例为 1/8 时, 效果最好, 识别准确率达到 84.4%, 比基线提升 0.4%. 因此, 在固定 VT-Shift 操作的移动比例为 1/8 的基础上, 增加 T-Conv 操作, 结果如图 3(b) 所示. 当卷积核大小为 7 时, 识别准确率进一步提升到 84.6%; 而当卷积核尺寸为 3 或 5 时, 则会导致识别准确率降低. 这说明当 T-Conv 采用较大的卷积核去关注远距离依赖时, 可以发挥与 VT-Shift 互补的时序建模作用. 短距离的 T-Conv 与 VT-Shift 重复对相邻帧之间的信息进行交换, 从而抵消了 VT-Shift 已经建立的时序关联, 导致识别准确率降低, 甚至减退到基线水平.

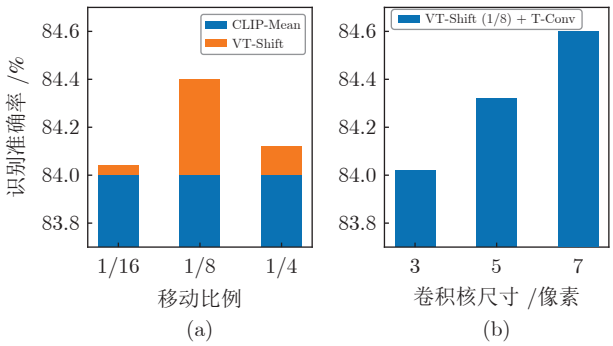

图 3 虚拟帧交互模块超参数消融实验结果
Fig. 3 Ablation studies of hyper-parameters of virtual frame interaction module

3.2.4 与其他提示学习方法对比

为进一步验证本文提出的视觉强化提示模块的有效性, 将本文模型与其他提示学习方法进行对比, 结果见表 3. 其中, ActionCLIP 和 CoOp 分别使用 16 个手工提示模板 (具体内容见附录 B) 和 16 个随机初始化的向量进行提示学习. 本文的视觉强化提示模块 VPM 获得了 85.7% 和 97.2% 准确率, 优于其他方法; 分类性能分别提升了 0.9% 和 0.2%, 验证了该模块的有效性. 这一提升依赖视觉强化提示模块融合视频类别和视觉分词所生成的视频自适应的语言提示, 初级语言表达经该提示强化后, 获得了更强的判别性.

3.3 全监督实验

本节使用完整的训练集和测试集分别进行训练和测试, 在 K400、UCF101 和 HMDB51 数据集上评估模型的性能. 模型训练过程中冻结语言编码器的参数, 只训练视频编码器和视觉强化提示模块. 表 4 为在 K400 数据集上, 本文模型与其他模型对比的实验结果, 其他模型包含基于 3D CNN 的模型[46−49]、基于 2D CNN 的模型[10−11, 50−51]、基于 ViT 的单模态模型[13−17, 22−23] 和基于语言−视觉对比学习的多模态模型[26−27] 共 4 个部分. 其中, CorrNet (Correlation network)[47]、SlowFast[48]、X3D-XXL[49] 和 MViT

表 3　提示学习方法的比较 (%)
Table 3　Comparisons of prompt learning methods (%)

方法	第 1 识别准确率	前 5 识别准确率
无	84.8	97.0
ActionCLIP	84.9	96.9
CoOp	85.5	97.1
本文模型	**85.7 (提升 0.9)**	**97.2 (提升 0.2)**

(B, 64 × 3) 从头进行训练; I3D NL[46] 和所有基于 2D CNN 的 TSM[10]、TEA (Temporal excitation and aggregation)[11]、TEINet (Temporal enhancement-and-interaction)[50]、TDN (Temporal difference network)[51] 在包含 1 000 个类别的 ImageNet 数据集上进行预训练; VTN (ViT-B)、TimeSformer (L)[15] 和 Swin (L) 在包含 21 000 个类别的 ImageNet 数据集上进行预训练; ViViT (L/16 × 2) 在数据规模为 3 亿的 JFT 数据集上进行预训练; EVL (ViT-B/16)、AIM (ViT-B/16) 和第 4、5 部分方法均是在数据规模为 4 亿的 WIT 数据集上进行预训练. 由表 4 可以看出, 基于 ViT 的模型整体上优于第 1、2 部分中基于 CNN 的模型, 显示了 ViT 在视频任务上的优势, 但同时也带来了更大的计算成本. 与之相比, 本文模型 (ViT-B/32) 仅用 8 帧输入就达到 80.5% 的第 1 识别准确率, 超过了第 1、2 部分中所有模型和使用更多输入帧的 VTN (ViT-B), 且每秒 10 亿次的浮点运算数 (Giga floating-point operations per second, GFLOPs) 更低, 彰显了多模态模型的巨大优势. 本文模型 (ViT-B/16) 的识别准确率超过第 3 部分所有基于 ViT 的单模态模型[13−17, 22−23], 同时也获得了表 4 所有模型中最高的识别准确率 84.1%, 且 GFLOPs 比 VTN (ViT-B)、ViViT (L) 和 TimeSformer (L) 分别降低了 29 倍、27 倍和 16 倍. 此外, 与其他多模态模型相比, 本文模型在所有设置下的实验结果均高于 Prom-

表 4　K400 数据集上, 全监督实验结果
Table 4　Fully-supervised experiment results on K400 dataset

类别	方法 (骨干网络)	预训练数据集	帧数	第 1 识别准确率 (%)	前 5 识别准确率 (%)	时间剪辑 × 空间裁剪	GFLOPs
3D CNN	I3D NL	ImageNet	32	77.7	93.3	10 × 3	359.0
	CorrNet	—	32	79.2	—	10 × 3	224.0
	SlowFast (R101-NL)	—	16 + 64	79.8	93.9	10 × 3	234.0
	X3D-XXL	—	16	80.4	94.6	10 × 3	144.0
2D CNN	TSM	ImageNet	16	74.7	91.4	10 × 3	65.0
	TEA	ImageNet	16	76.1	92.5	10 × 3	70.0
	TEINet	ImageNet	16	76.2	92.5	10 × 3	66.0
	TDN	ImageNet	8 + 16	79.4	93.9	10 × 3	198.0
ViT	VTN (ViT-B)	ImageNet	250	78.6	93.7	1 × 1	4218.0
	ViViT (L/16 × 2)	JFT	32	83.5	95.5	4 × 3	3992.0
	TimeSformer (L)	ImageNet	96	80.7	94.7	1 × 3	2380.0
	MViT (B, 64 × 3)	—	64	81.2	95.1	3 × 3	455.0
	Swin (L)	ImageNet	32	83.1	95.9	4 × 3	604.0
	EVL (ViT-B/16)	WIT	8	82.9	—	3 × 1	444.0
	AIM (ViT-B/16)	WIT	8	83.9	96.3	3 × 1	606.0
语言–视觉对比学习	PromptCLIP (A6)	WIT	16	76.9	93.5	5 × 1	—
	ActionCLIP (ViT-B/32)	WIT	8	78.4	94.3	1 × 1	35.0
	ActionCLIP (ViT-B/16)	WIT	8	81.1	95.5	1 × 1	141.0
	ActionCLIP (ViT-B/16)	WIT	16	82.6	96.2	10 × 3	282.0
	ActionCLIP (ViT-B/16)	WIT	32	83.8	**97.1**	10 × 3	563.0
	X-CLIP (ViT-B/32)	WIT	8	80.4	95.0	4 × 3	39.0
	X-CLIP (ViT-B/32)	WIT	16	81.1	95.5	4 × 3	75.0
	X-CLIP (ViT-B/16)	WIT	8	83.8	96.7	4 × 3	145.0
本文模型	本文模型 (ViT-B/32)	WIT	8	80.5	95.1	4 × 3	39.8
	本文模型 (ViT-B/32)	WIT	16	81.4	95.5	4 × 3	75.6
	本文模型 (ViT-B/32)	WIT	32	83.1	95.7	4 × 3	144.2
	本文模型 (ViT-B/16)	WIT	8	**84.1**	96.7	4 × 3	145.8

ptCLIP; 在相同基础网络和采样帧数下, 本文模型较 ActionCLIP 具有明显性能优势, 并且本文模型 (ViT-B/16) 在 8 帧输入设置下的识别准确率高于使用 32 帧输入的 ActionCLIP (ViT-B/16); 与当前先进模型 X-CLIP 相比, 本文模型仅增加较小的计算量, 就获得了相同基础网络和输入帧设置下的更优性能.

图 4 直观地展示了表 4 中基于 ViT 的模型及其变体、多模态模型及其变体和本文模型的变体在识别准确率和计算成本方面的平衡情况, 横轴表示 GFLOPs, 纵轴表示第 1 识别准确率. 由图 4 可以看出, AIM 虽然可以获得更优的识别准确率, 但 GFLOPs 也显著增加. 除 AIM 外, 基于 ViT 的单模态模型的识别准确率折线整体处于本文模型的右下方, 体现了虚拟帧交互模块轻量级时序建模的优势和本文模型引入语言监督对提升视频表达质量和分类性能的有效性. 与 ActionCLIP 相比, 本文模型以更小计算成本显著超越其性能; 与 X-CLIP 相比, 本文模型以可忽略的 GFLOPs 增加换取了性能上的提升.

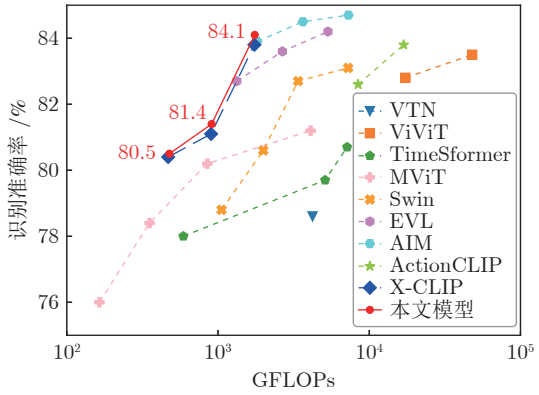

图 4 在 K400 数据集上, 本文模型与目前先进模型的识别准确率和浮点运算数比较

Fig. 4 Comparisons of accuracy and GFLOPs between the proposed model and state-of-the-art models on K400 dataset

表 5 总结了在 HMDB51 和 UCF101 数据集上, 基于 2D CNN[45, 52–54]、3D CNN[55–57] 和多模态对比学习的视频行为识别模型和本文模型的分类性能. 所有模型都先在 K400 上进行预训练, 再在 UCF101 和 HMDB51 数据集上进行微调. 本文模型 (ViT-B/16) 使用 8 帧输入, 在 HMDB51 和 UCF101 数据集上分别获得 76.7% 和 97.6% 的第 1 识别准确率, 与其他模型相比, 获得了最佳性能.

表 5 总结了 TSN、TBN (Temporal Bilinear Network)[52]、PPAC (Pyramid pyramid attention clusters)[53]、TCP (Temporal-attentive covariance pooling)[54] 等基于 2D CNN 的模型, ARTNet (Appearance and relation network)[55]、R3D (3D ResNet)[56]、MCL (Motion-focused contrastive learning)[57] 等基于 3D CNN 的模型以及多模态对比学习模型和本文模型在 UCF101 和 HMDB51 数据集上的分类性能. 由全监督实验结果可以看出, 本文模型与目前流行的基于 ViT 的模型和多模态模型相比, 都取得了十分有竞争力的性能. 这归因于 2 个因素: 1) 虚拟帧交互模块能有效地在 ViT 结构的中间层建模视频的时间依赖; 2) 本文提出的模型可以有效适应视频行为识别任务, 展现其强大的视频表达能力.

表 5 HMDB51 和 UCF101 数据集上, 全监督实验结果

Table 5 Fully-supervised experiment results on HMDB51 and UCF101 datasets

方法 (骨干网络)	帧数	UCF101 (%)	HMDB51 (%)
TSN (2D R50)	8	91.7	64.7
TBN (2D R34)	8	93.6	69.4
PPAC (2D R152)	20	94.9	69.8
TCP (2D TSN R50)	8	95.1	72.5
ARTNet (3D R18)	16	94.3	70.9
R3D (3D R50)	16	92.9	69.4
MCL (R (2 + 1) D)	16	93.4	69.1
ActionCLIP (ViT-B/16)	32	97.1	76.2
X-CLIP (ViT-B/32)	8	95.3	72.8
X-CLIP (ViT-B/16)	8	97.4	75.6
本文模型 (ViT-B/32)	8	96.1	74.3
本文模型 (ViT-B/16)	8	**97.6**	**76.7**

3.4 小样本实验

遵循一般设置, 从 UCF101 和 HMDB51 数据集的每个类别中, 随机抽取 $K = \{2, 4, 8, 16\}$ 个视频构建训练集, 使用第 1 种划分方式进行评估. 实验使用 8 帧输入的本文模型 (ViT-B/16) 进行单视图测试. 表 6 为本文模型在小样本实验上的第 1 识别准确率, 并与其他表现优秀的单模态和多模态模型对比. 表 6 中, 单模态模型 TSM、TimeSformer 和 Swin (B) 与多模态模型 ActionCLIP (ViT-B/16)、X-CLIP (ViT-B/16) 相比, 存在显著的性能差距, 在 $K = 2$ 的极端情况下, 差距最为明显, 这说明单模态模型由于数据的严重缺乏, 发生了过拟合.

随着数据量增加, 过拟合问题得到缓解. 本文模型在相同数据条件下, 显示出明显的性能优势和稳健性. 在 $K = 2$ 时, 本文模型 (ViT-B/16) 在 HMDB51 和 UCF101 数据集上的第 1 识别准确率分别超过 Swin (B) 28.7% 和 23.1%. 这得益于语言信息对视觉分支的监督作用, 也进一步验证了本文模型将 CLIP 的知识迁移到视频小样本识别的有效性.

表 6　HMDB51 和 UCF101 数据集上, 小样本实验结果 (%)
Table 6　Few-shot experiment results on HMDB51 and UCF101 datasets (%)

方法 (骨干网络)	HMDB51				UCF101			
	$K=2$	$K=4$	$K=8$	$K=16$	$K=2$	$K=4$	$K=8$	$K=16$
TSM	17.5	20.9	18.4	31.0	25.3	47.0	64.4	61.0
TimeSformer	19.6	40.6	49.4	55.4	48.5	75.6	83.7	89.4
Swin (B)	20.9	41.3	47.9	56.1	53.3	74.1	85.8	88.7
ActionCLIP (ViT-B/16)	43.7	51.2	55.6	64.2	73.7	80.2	86.3	89.8
X-CLIP (ViT-B/16)	49.5	54.6	57.7	65.3	76.3	81.4	85.9	89.4
本文模型 (ViT-B/16)	**49.6**	**54.9**	**58.8**	**65.5**	**76.4**	**82.1**	**86.7**	**90.1**

同时, 与 ActionCLIP 和 X-CLIP 相比, 在提供完全相同的训练数据 (见附录 C) 时, 本文模型表现出最优性能.

3.5　零样本实验

零样本实验十分具有挑战性, 要求模型对从未训练过的类别进行测试. 与文献 [18] 相同, 首先使用 8 帧输入的本文模型 (ViT-B/16) 在 K400 上进行预训练, 然后分别在 UCF101 和 HMDB51 的 3 种划分的测试集上进行测试, 表 7 为 3 种划分的第 1 识别准确率的平均值.

表 7　HMDB51 和 UCF101 数据集上, 零样本实验结果 (%)
Table 7　Zero-shot experiment results on HMDB51 and UCF101 datasets (%)

方法 (骨干网络)	HMDB51	UCF101
ZSECOC	22.6	15.1
UR	24.4	17.5
TS-GCN	23.2	34.2
E2E	32.7	48.0
ER-ZSAR	35.3	51.8
ActionCLIP	41.9	66.6
X-CLIP (ViT-B/16)	43.5	70.9
本文模型 (ViT-B/16)	**44.0**	**72.6**

如表 7 所示, 对比单模态模型 ZSECOC (Zero-shot error-correcting output codes)[58]、UR (Universal representation)[59]、TS-GCN (Two-stream graph convolutional network)[60]、E2E (End-to-end algorithm for zero-shot learning in video classification)[61] 和 ER-ZSAR (Elaborative rehearsal enhanced zero-shot action recognition)[62], 本文模型在 UCF101 和 HMDB51 上, 分别比以往最佳结果 ER-ZSAR 提升了 20.8% 和 8.7%. 对比多模态模型, 使用完全相同的训练方式, 本文模型在 HMDB51 和 UCF101 上的识别准确率分别优于 ActionCLIP 2.1% 和 6.0%, 优于 X-CLIP 0.5% 和 1.7%. 这种显著改进说明本文模型可以适应视频行为识别任务, 获得了泛化性能强大的视频表达.

4　结束语

本文通过对 CLIP 模型进行扩展, 提出一种适用于视频行为识别任务的多模态模型. 该模型在视觉编码器中构造了虚拟帧交互模块, 完成网络中间层的跨帧信息交互, 更好地捕获了视频远距离及相邻帧之间的时间依赖信息; 同时, 在语言分支上构建了视觉强化提示模块, 通过注意力机制融合视觉分支输出分词中包含的视觉信息, 自动生成适应视频行为识别的语言提示, 强化视频的语言表达. 在全监督、小样本和零样本 3 种不同实验场景下的实验结果表明了本文多模态模型在视频行为识别任务上的有效性和泛化性.

在未来工作中, 将模型拓展到不同的视频理解任务中, 例如视频检索、视频标注、视频语义分割等. 此外, 考虑构建更优的语言提示学习模块和时空建模方式, 以进一步增强多模态模型的识别准确率.

附录 A　模型训练超参数

3 种实验设置下模型训练超参数如表 A1 所示, 其中设置随机初始化参数的学习率比基础学习率高 10 倍.

附录 B　手工提示模板

第 3.2.4 节中, 为将视觉强化提示模块与手工提示模板进行对比, 构建如下 16 个手工模板:

A photo of action {label}; A picture of action {label}; Human action of {label}; {label}, an action; {label}, this is an action; {label}, a video of action; Playing action of {label}; {label}; Playing a kind of action, {label}; Doing a kind of action, {label}; Look, the human is {label}; Can you recognize the action of {label}?; Video classification of {label}; A video of {label}; The man is {label}; The woman is {label}.

表 A1 模型训练超参数
Table A1 Hyper-parameters for model training

超参数	全监督	小样本	零样本
VPM α	0.1	0.1	0.1
VPM β	0.1	0.1	0.1
优化器	AdamW	AdamW	—
优化器 β 值	(0.90, 0.98)	(0.90, 0.98)	—
批大小	256	64	256
学习率策略	cosine	cosine	—
预热轮数	5	5	—
基础学习率	8×10^{-6}	2×10^{-6}	—
最小学习率	8×10^{-8}	2×10^{-8}	—
轮数	30	50	—
随机翻转	0.5	0.5	0.5
多尺度裁剪	(1, 0.875, 0.75, 0.66)	(1, 0.875, 0.75, 0.66)	(1, 0.875, 0.75, 0.66)
颜色抖动	0.8	0.8	0.8
灰度值	0.2	0.2	0.2
标签平滑	0.1	0.1	0.1
混合	0.8	0.8	0.8
切割混合	1.0	1.0	1.0
权重衰减	0.001	0.001	0.001

附录 C 小样本实验

第 3.4 节对模型在小样本实验设置下的性能进行评估. 表 6 中 TSM、TimeSformer 和 Swin (B) 的实验结果引自 X-CLIP, 训练和测试使用原论文默认超参数训练.

为进行公平对比, ActionCLIP、X-CLIP 和本文模型使用相同 K 个随机样本进行训练, 随机样本在 UCF101 和 HMDB51 数据集上第 1 种划分中随机抽取得到. 表 C1 展示了小样本实验所使用的随机样本编号.

表 C1 小样本实验使用的随机样本
Table C1 Random examples used in few-shot experiment

数据集	K 值	随机样本在第 1 种划分中的编号
HMDB51	2	[22, 25]
	4	[69, 9, 21, 36]
	8	[44, 47, 64, 67, 69, 9, 21, 6]
	16	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
UCF101	2	[21, 48]
	4	[2, 16, 23, 44]
	8	[14, 20, 60, 27, 33, 9, 21, 32]
	16	[14, 20, 60, 27, 33, 9, 21, 32, 8, 15, 1, 26, 38, 44, 60, 48]

References

1 Zhou Bo, Li Jun-Feng. Human action recognition combined with object detection. *Acta Automatica Sinica*, 2020, **46**(9): 1961−1970
(周波, 李俊峰. 结合目标检测的人体行为识别. 自动化学报, 2020, **46**(9): 1961−1970)

2 Yang Tian-Jin, Hou Zhen-Jie, Li Xing, Liang Jiu-Zhen, Huan Juan, Zheng Ji-Xiang. Recognizing action using multi-center subspace learning-based spatial-temporal information fusion. *Acta Automatica Sinica*, 2022, **48**(11): 2823−2835
(杨天金, 侯振杰, 李兴, 梁久祯, 宦娟, 郑纪翔. 多聚点子空间下的时空信息融合及其在行为识别中的应用. 自动化学报, 2022, **48**(11): 2823−2835)

3 Zuo Guo-Yu, Xu Zhao-Kun, Lu Jia-Hao, Gong Dao-Xiong. A structure-optimized DDAG-SVM action recognition method for upper limb rehabilitation training. *Acta Automatica Sinica*, 2020, **46**(3): 549−561
(左国玉, 徐兆坤, 卢佳豪, 龚道雄. 基于结构优化的 DDAG-SVM 上肢康复训练动作识别方法. 自动化学报, 2020, **46**(3): 549−561)

4 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: 2014. 568−576

5 Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 4489−4497

6 Carreira J, Zisserman A. Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 4724−4733

7 Qiu Z, Yao T, Mei T. Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017. 5534−5542

8 Du T, Wang H, Torresani L, Ray J, Paluri M. A closer look at spatio-temporal convolutions for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6450−6459

9 Xie S, Sun C, Huang J, Tu Z, Murphy K. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision. Munich, Germany: Springer, 2018. 318−335

10 Lin J, Gan C, Han S. TSM: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 7082−7092

11 Li Y, Ji B, Shi X T, Zhang J G, Kang B, Wang L M. TEA: Temporal excitation and aggregation for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 909−918

12 Liu Z Y, Wang L M, Wu W, Qian C, Lu T. TAM: Temporal adaptive module for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 13688−13698

13 Neimark D, Bar O, Zohar M, Asselmann D. Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 3156−3165

14 Arnab A, Dehghani M, Heigold G, Sun C, Lucic M, Schmid C. ViViT: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 6816−6826

15 Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR, 2021. 813−824

16 Fan H Q, Xiong B, Mangalam K, Li Y H, Yan Z C, Malik J, et al. Multi-scale vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 6824−6835

17 Liu Z, Ning J, Cao Y, Wei Y X, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 3192−3201

18 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR, 2021. 8748−8763

19 Jia C, Yang Y F, Xia Y, Chen Y T, Parekh Z, Pham H, et al. Scaling up visual and vision-language representation learning with noisy text supervision. In: Proceedings of the 38th International Conference on Machine Learning. Vienna, Austria: PMLR, 2021. 4904−4916

20 Yuan L, Chen D D, Chen Y L, Codella N, Dai X Y, Gao J F, et al. Florence: A new foundation model for computer vision. arXiv preprint arXiv: 2111.11432, 2021.

21 Pan J T, Lin Z Y, Zhu X T, Shao J, Li H S. ST-Adapter: Parameter-efficient image-to-video transfer learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: MIT Press, 2022. 1−16

22 Lin Z Y, Geng S J, Zhang R R, Gao P, Melo G D, Wang X G, et al. Frozen CLIP models are efficient video learners. In: Proceedings of the European Conference on Computer Vision. Tel-Aviv, Israel: Springer, 2022. 388−404

23 Yang T J N, Zhu Y, Xie Y S, Zhang A, Chen C, Li M. AIM: Adapting image models for efficient video action recognition. In: Proceedings of the International Conference on Learning Representations. Kigali, Republic of Rwanda: 2023. 1−18

24 Xu H, Ghosh G, Huang P Y, Okhonko D, Aghajanyan A, Metze F, et al. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021. 6787−6800

25 Ju C, Han T D, Zheng K H, Zhang Y, Xie W D. Prompting visual-language models for efficient video understanding. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 105−124

26 Wang M M, Xing J Z, Liu Y. ActionCLIP: A new paradigm for video action recognition. arXiv preprint arXiv: 2109.08472, 2021.

27 Ni B L, Peng H W, Chen M H, Zhang S Y, Meng G F, Fu J L, et al. Expanding language-image pretrained models for general video recognition. In: Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel: Springer, 2022. 1−18

28 Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, USA: 2017. 6000−6010

29 Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X H, Unterthiner T, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. Vienna, Austria: 2021. 1−14

30 Brown T B, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Red Hook, USA: 2020. 1877−1901

31 Gao T Y, Fisch A, Chen D Q. Making pre-trained language models better few-shot learners. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Bangkok, Thailand: ACL, 2021. 3816−3830

32 Jiang Z B, Xu F F, Araki J, Neubig G. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 2020, **8**: 423−438

33 Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Virtual Event: ACL, 2021. 255−269

34 Shin T, Razeghi Y, Logan IV R L, Wallace E, Singh S. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Virtual Event: ACL, 2020. 4222−4235

35 Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Punta Cana, Dominican Republic: ACL, 2021. 3045−3059

36 Li X L, Liang P. Prefix-tuning: Optimizing continuous prompts for generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Virtual Event: ACL, 2021. 4582−4597

37 Zhou K Y, Yang J K, Loy C C, Liu Z W. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 2022, **130**(9): 2337−2348

38 Zhou K Y, Yang J K, Loy C C, Liu Z W. Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE, 2022. 16795−16804

39 Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with sub-word units. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany: ACL, 2016. 1715−1725

40 Zhang H, Hao Y B, Ngo C W. Token shift transformer for video classification. In: Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM Press, 2021. 917−925

41 Xie J T, Zeng R R, Wang Q L, Zhou Z Q, Li P H. SoT: Delving deeper into classification head for transformer. arXiv preprint arXiv: 2104.10935, 2021.

42 Kay W, Carreira J, Simonyan K, Zhang B, Hilier C, Vijayanarasimhan S, et al. The kinetics human action video dataset. arXiv preprint arXiv: 1705.06950, 2017.

43 Soomro K, Zamir A R, Shah M. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv: 1212.0402, 2012.

44 Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T. HMDB: A large video database for human motion recognition. In: Proceedings of the International Conference on Computer Vision. Barcelona, Spain: IEEE, 2011. 2556−2563

45 Wang L M, Xiong Y J, Wang Z, Qiao Y, Lin D H, Tang X O, et al. Temporal segment networks: Towards good practices for deep action recognition. In: Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016. 20−36

46 Wang X L, Girshick R, Gupta A, He K M. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 7794−7803

47 Wang H, Tran D, Torresani L, Feiszli M. Video modeling with correlation networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 349−358

48 Feichtenhofer C, Fan H Q, Malik J, He K M. SlowFast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE, 2019. 6201−6210

49 Feichtenhofer C. X3D: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 203−213

50 Liu Z Y, Luo D H, Wang Y B, Wang L M, Tai Y, Wang C J, et al. TEINet: Towards an efficient architecture for video recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**(7): 11669−11676

51 Wang L M, Tong Z, Ji B, Wu G S. TDN: Temporal difference networks for efficient action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE, 2021. 1895−1904

52 Li Y H, Song S J, Li Y Q, Liu J Y. Temporal bilinear networks

for video action recognition. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence and the 31th Innovative Applications of Artificial Intelligence Conference and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence. Hawaii, USA: AAAI, 2019. 8674−8681

53 Long X, Melo G D, He D L, Li F, Chi Z Z, Wen S L, et al. Purely attention based local feature integration for video classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(4): 2140−2154

54 Gao Z L, Wang Q L, Zhang B B, Hu Q H, Li P H. Temporal-attentive covariance pooling networks for video recognition. *Advances in Neural Information Processing Systems*, 2021, **34**: 13587−13598

55 Wang L M, Li W, Li W, Gool L V. Appearance-and-relation networks for video classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 1430−1439

56 Kataoka H, Wakamiya T, Hara K, Satoh Y. Would mega-scale datasets further enhance spatio-temporal 3D CNNs? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 6546−6555

57 Li R, Zhang Y H, Qiu Z F, Yao T, Liu D, Mei T. Motion-focused contrastive learning of video representations. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 2105−2114

58 Qin J, Liu L, Shao L, Shen F, Ni B B, Chen J X, et al. Zero-shot action recognition with error-correcting output codes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE, 2017. 1042−1051

59 Zhu Y, Long Y, Guan Y, Newsam S, Shao L. Towards universal representation for unseen action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018. 9436−9445

60 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, **33**(1): 8303−8311

61 Brattoli B, Tighe J, Zhdanov F, Perona P, Chalupka K. Rethinking zero-shot video classification: End-to-end training for realistic applications. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020. 4613−4623

62 Chen S Z, Huang D. Elaborative rehearsal for zero-shot action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE, 2021. 13638−13647

张 颖 大连民族大学计算机科学与工程学院硕士研究生. 主要研究方向为视频行为识别.
E-mail: z_ying1201@126.com
(**ZHANG Ying** Master student at the School of Computer Science and Engineering, Dalian Minzu University. Her main research interest is video action recognition.)

张冰冰 大连理工大学电子信息与电气工程学部博士研究生. 2016 年获得长春工业大学硕士学位. 主要研究方向为人体行为识别, 图像分类和深度学习.
E-mail: icyzhang@mail.dlut.edu.cn
(**ZHANG Bing-Bing** Ph.D. candidate in the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. She received her master degree from Changchun University of Technology in 2016. Her research interest covers human action recognition, image classification, and deep learning.)

董 微 大连民族大学计算机科学与工程学院硕士研究生. 主要研究方向为视频行为识别.
E-mail: vvvDongWei@163.com
(**DONG Wei** Master student at the School of Computer Science and Engineering, Dalian Minzu University. Her main research interest is video action recognition.)

安峰民 大连民族大学计算机科学与工程学院硕士研究生. 主要研究方向为视频行为识别.
E-mail: anfengmin@163.com
(**AN Feng-Min** Master student at the School of Computer Science and Engineering, Dalian Minzu University. His main research interest is video action recognition.)

张建新 大连民族大学计算机科学与工程学院教授. 主要研究方向为计算机视觉, 智能医学影像分析. 本文通信作者.
E-mail: jxzhang0411@163.com
(**ZHANG Jian-Xin** Professor at the School of Computer Science and Engineering, Dalian Minzu University. His research interest covers computer vision and intelligent medical image analysis. Corresponding author of this paper.)

张 强 大连理工大学电子信息与电气工程学部教授. 主要研究方向为大数据分析与处理, 机器行为与人机协同, 生物计算和人工智能.
E-mail: zhangq@dlut.edu.cn
(**ZHANG Qiang** Professor in the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology. His research interest covers big data analysis and processing, machine behavior and human machine collaboration, bio-computing, and artificial intelligence.)