基于增强负例多粒度区分模型的视频动作识别研究

刘良振[1]，杨阳[1]，夏莹杰[2,3]，邝砾[1]
（1. 中南大学计算机学院，湖南 长沙 410083；2. 杭州电子科技大学微电子研究院，浙江 杭州 310005；
3. 浙江大学计算机科学与技术学院，浙江 杭州 310012）

摘　要：为提升模型对视频动作的细粒度区分能力，提出一种基于对比学习的增强负例区分范式。通过为每个视频生成增强负例集合，以补充最难区分的视频-文本负例对。为了进一步区分正负例，基于该范式提出一种用于视频动作识别的多粒度区分模型。在该模型中，视频表征器通过引入文本正例特征引导视频特征提取，而正负语义区分器利用自注意力机制构建正负语义之间的自相关关系。该模型既能够实现模态间视频与增强负例集的粗粒度区分，还可以实现文本模态内正例与增强负例集的细粒度区分。实验结果表明，增强负例集能显著提升模型在细粒度类别标签上的识别能力，多粒度区分模型在Kinetics-400、HMDB51和UCF101数据集上的性能均优于当前较具代表性的方法。

关键词：对比学习；增强负例；范式；视频动作识别
中图分类号：TP391.41
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024268

Study on video action recognition based on augment negative example multi-granularity discrimination model

LIU Liangzhen[1], YANG Yang[1], XIA Yingjie[2,3], KUANG Li[1]
1. School of Computer Science and Engineering, Central South University, Changsha 410083, China
2. Micro-Electronics Research Institute, Hangzhou Dianzi University, Hangzhou 310005, China
3. College of Computer Science and Technology, Zhejiang University, Hangzhou 310012, China

Abstract: An augment negative example discrimination paradigm based on contrastive learning was proposed to improve the model's fine-grained discrimination ability of video actions. The most challenging video-text negative pairs was generated, forming an augmented negative example set for each video sample. Based on this paradigm, a multi-granularity discrimination model for video action recognition was proposed to further distinguish between positive and negative examples. In this model, video features were extracted by the video representation module guided by textual positive examples, while self-correlation relationships between positive and negative semantics were established by the semantic discriminator equipped with a self-attention mechanism. Meanwhile, a coarse-grained distinction between the video modality and the augmented negative example set was achieved, while a fine-grained distinction between positive examples and the augmented negative example set within the text modality was also accomplished. Experimental results demonstrate that the augment negative set improves the model's recognition ability on fine-grained class labels, and the multi-granularity discrimination model outperforms current state-of-the-art methods on the Kinetics-400, HMDB51 and UCF101 datasets.

Keywords: contrastive learning, augmented negative examples, paradigm, video action recognition

收稿日期：2024–07–29；修回日期：2024–12–05
通信作者：邝砾，kuangli@csu.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2022YFF0902500）
Foundation Item：The National Key Research and Development Program of China (No.2022YFF0902500)

0　引言

视频是一种常见的媒体形式[1]，也是人们获取信息的重要渠道。视频动作识别旨在对视频场景中的动作或行为进行分类，是实现视频内容理解的重要任务之一[2]，在安全监控、医疗保健、人机交互、智能交通和体育分析等领域具有广泛的应用前景[3-5]。

在视频动作识别技术的发展过程中，基于深度学习的模型逐渐占据主导地位。DeepVideo[6]利用视频的局部时空信息建模，是将卷积神经网络（CNN, convolutional neural network）应用于视频分类的方法之一。随后，双流网络被引入视频动作识别任务中，文献[7]将视频的光流特征和深度特征结合，用于学习视频的时序信息，从而形成融合深度特征和光流特征的双流网络。为了减少单独提取光流特征的计算开销，基于3D卷积的方法开始发展，文献[8]利用3D卷积核直接从视频中提取一体化的时空特征。然而，基于3D卷积的视频动作识别模型存在参数量庞大、极度依赖计算资源的问题，促使学者们转向研究基于高效视频建模的方法。文献[9]提出一种沿时间维度移动部分通道的模块，能够促进相邻帧之间的信息交换，此模块可以嵌入2D卷积神经网络中，实现了零计算开销和零参数的时间建模。

然而，传统的视频动作识别方法通常执行一对多类的投票任务，未考虑类别文本的语义信息，视觉—语言模型（VLM, visual-language model）的发展有效弥补了这一缺陷。VLM的图像编码器可以捕捉深度视觉特征，文本编码器能将文本转化为语义向量表示，进而通过对齐视觉和语言模态特征构建2个模态间的桥梁。VLM通过在大规模图像—文本对上进行预训练，使得其在视觉任务中具备强大的知识可迁移性。将VLM应用于视频动作识别领域，不仅可以充分利用大规模图像—文本对中对齐的先验知识，还可以在视频时序建模中实现视频与文本的深度特征对齐。因此，基于VLM的视频动作识别方法具有显著的性能提升潜力。对比语言图像预训练（CLIP, contrastive language-image pre-training）[10]模型是一种当前主流的VLM，该模型通过在4亿个图像-文本对上进行对比学习，实现视觉和语言对齐。文献[11]将CLIP引入视频动作识别任务，在Kinetics-400数据集[12]上取得82.6%的Top-1准确率，突破了传统方法的性能瓶颈。然而，在对比学习过程中，正负例对的采样是随机且无方向性的，这可能导致模型遗漏一些最难区分的负例对，从而削弱了模型对语义相似类别的识别能力。因此，如何为模型补充最难区分的负例对，成为提升模型性能的关键问题。

Cai等[13]指出，在对比学习中，95%的简单负例对模型的影响非常小，而其余5%最难区分的负例对模型性能具有决定性作用。例如，在Kinetics-400数据集[12]中，“dunking basketball”为扣篮动作的类别标签，而在该类视频中扣篮之前往往还包含其他相关的篮球动作，如运球（dribbling basketball）、投篮（shooting basketball）等，这些与视频相关的语义类别容易导致模型进行视频动作识别时出现混淆。对于“dunking basketball”而言，“dribbling basketball”和“shooting basketball”是其最难区分的负例类别。因此，强化视频与其相关语义间的细粒度区分至关重要。

如图1(a)所示，传统范式以视频与其对应标签作为正例对，以同批次视频与其他类别标签作为负例对进行模型对比训练。然而，随机采样的同批次视频-文本对未考虑模型最难区分的负例对，可能削弱模型对相似类别的识别能力。如图1(b)所示，本文提出一种增强负例区分范式，旨在强化视频与其相关语义类别的区分度。首先，通过为每个视频-文本正例对补充最难区分的负例对集合，为对比学习提供了方向性更明确的负例对，这是在负例层面的增强策略。然后，将该策略扩展融合到相似度矩阵计算中，从而增强视频与其相关语义类别的区分度。

总体而言，目前基于VLM的视频动作识别方法仍面临以下挑战。

1) 视频动作识别任务样例中，通常存在一系列语义相似的场景或行为，这些与视频相关的类别容易导致模型在视频动作分类时出现误判。此外，随机采样的一个批次正负例样本对具有随机性且无方向性，可能遗漏最难区分的负例对，从而削弱模型对语义相似类别的识别能力。

2) 当前模型往往只考虑模态层面视频-文本的粗粒度对齐，而忽略了视频-文本表征区分以及语义层面相关类别表征区分，从多粒度角度提升模型的区分表征能力是当前亟待解决的问题之一。

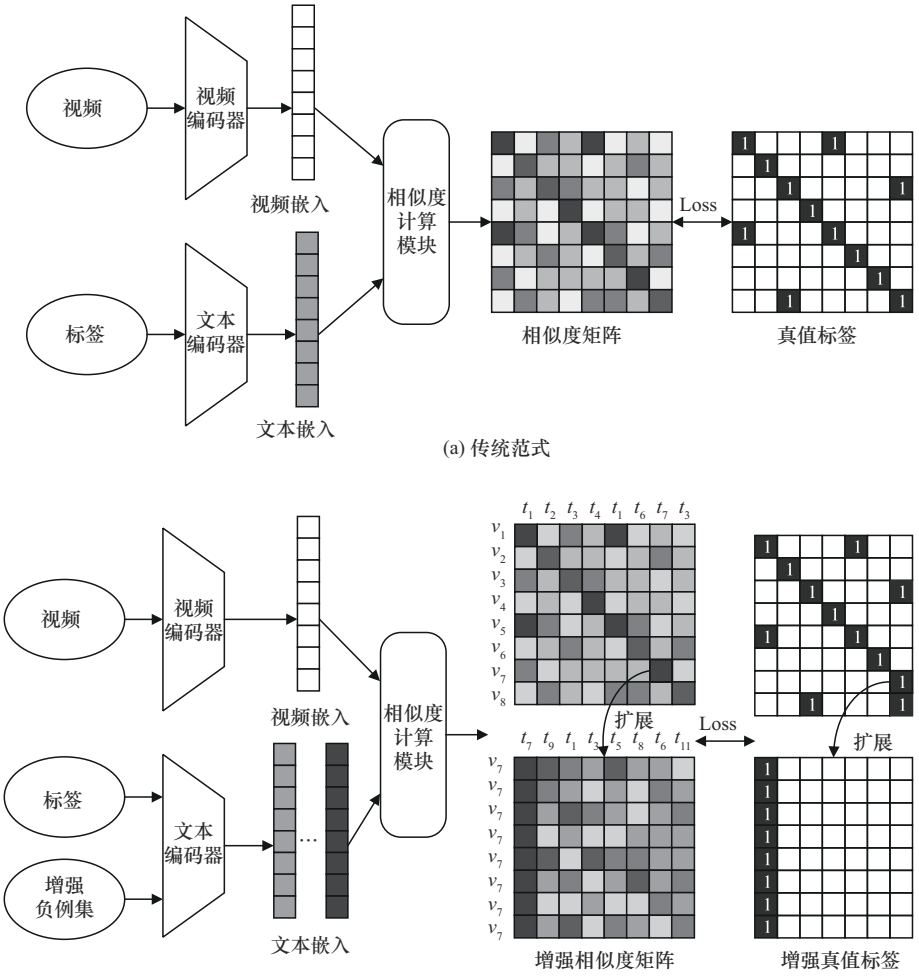


图1　传统范式与增强负例范式对比

为了解决上述挑战，本文主要的研究工作如下。

1) 提出一种增强负例区分范式，通过离线检索每个视频样例的增强负例集合，为模型补充最难区分的视频-文本负例对，从而提升模型对视频与其相关语义类别的区分能力。此外，本文设计了一种与该范式适配的增强负例集生成方式，基于CLIP卓越的视觉-语言特征对齐能力，精准定位与视频强相关的负例。

2) 基于增强负例区分范式提出一种用于视频动作识别的多粒度区分模型（MGDM, multi-granularity discrimination model），该模型包括视频表征器和正负语义区分器。视频表征器引入文本正例特征以引导帧间时序建模，正负语义区分器采用自注意力机制构建正负语义间的自相关关系。该模型不仅关注模态间视频与增强负例集的粗粒度区分，还关注文本模态内正例与增强负例集的细粒度区分。

3) 在Kinetics-400、HMDB51和UCF101数据集上对基于增强负例区分范式的多粒度区分模型进行测试，实验结果表明，本文方法的性能优于当前较具代表性的视频动作识别方法。

1　相关工作

在视频动作识别领域，随着深度学习技术的不断进步，各种先进方法不断被提出，以提升识别的准确性和效率。根据视频动作识别技术的发展进程，本文将基于深度学习的主要方法分为4类，包括基于双流网络的方法、基于3D卷积的方法、基于高效视频建模的方法和基于VLM的方法。

1.1　基于双流网络的方法

基于双流网络的视频动作识别方法利用光流[14]的特点，有效去除了不运动背景的干扰，从

而表征帧间的时间关系。文献[7]方法的提出引发了学者们对双流网络的广泛关注，Wang等[15]提出轨迹池化深度卷积描述符（TDD, trajectory-pooled deep-convolutional descriptor），用于视频动作识别，与手工设计的特征相比，该方法具有更高的判别能力。Donahue等[16]提出长期循环卷积网络（LRCN, long-term recurrent convolutional network），该网络可以将可变长视频帧数直接映射为可变长自然语言文本输出，并对复杂的时间动态进行建模。Wang等[17]基于长时间结构建模的思想，结合稀疏时间采样策略和视频级监督，提出一种时序分割网络（TSN, temporal segment network）用于视频动作识别。为捕获长期依赖关系，Feichtenhofer等[18]利用动态门控机制将双流架构中的外观流和路径流相结合，提出一种基于时空特征乘性交互的通用ConvNet结构，实现端到端的视频动作识别。这类方法虽然在视频动作识别中表现出色，但光流特征的提取需要耗费大量计算资源，且光流特征与深度特征的提取不同步，难以提取一体化的视频特征。

1.2　基于3D卷积的方法

基于视频的空间和时间特性，3D卷积能够同步建模视频的时空特征，具有良好的适配性。文献[19-20]都将3D卷积应用于视频动作识别领域，实现了基础性探索。Hara等[21]研究了现有视频数据集是否足够支持训练具有时空3D卷积核的极深卷积神经网络。为应对基于3D卷积神经网络在视频动作识别中面临的空间、时间和计算复杂性等挑战，Xie等[22]建立了一个有效的视频分类系统，在计算速度和分类准确性之间寻求平衡。文献[23]提出一种将视频识别拆分为慢速和快速2个分支的架构，其中慢速率的分支捕获空间背景特征，快速率的分支捕获时间线中的运动特征。在文献[23]的基础上，Feichtenhofer[24]进一步提出X3D网络，X3D采用逐步网络扩展方式，在每步中扩展一个轴，以实现良好的精度与复杂性之间的平衡。实验结果表明，高时空分辨率的X3D网络表现良好。尽管这类方法能够提取一体化的视频特征，但由于参数量庞大，仍面临计算资源需求巨大的挑战。

1.3　基于高效视频建模的方法

随着数据集规模和部署需求的增加，基于高效视频建模的视频动作识别方法逐渐涌现[25]。TSM[9]是这类方法中具有代表性的工作之一，该方法在零新增参数与零计算的情况下，对视频相邻帧间的时序关系进行建模。为了使视频动作方法应用于实时性要求较高的场景，Piergiovanni等[26]提出一种微小视频网络（TVN, tiny video network），该网络为实时视频应用提供了新工具。Jiang等[27]提出一个时空模块（STM, spatio temporal module），该模块由通道时空模块和通道运动模块组成，通过有限的额外计算显著提升视频任务的性能。时间激励和聚合（TEA, temporal excitation and aggregation）模块[28]中包含运动激励模块和多重时间聚合模块，用于捕捉短时间和长时间的动态变化。为了捕捉视频不同的运动模式，Liu等[29]提出一种时间自适应模块（TAM, temporal adaptive module），该模块可以集成到2D卷积神经网络中，仅需增加极少的计算开销。Wang等[30]提出时间差分网络（TDN, temporal difference network），其核心是捕捉多尺度的空间信息以用于视频动作识别任务。针对传统方法只采样少量帧的问题，Zhang等[31]提出放大和聚焦网络（AFNet, ample and focal network）通过中间特征中的动态选择强制执行隐式时间建模，显著提升了模型效率。这类方法通过优化计算实现了较高的效率，但在视频动作识别准确率上难以实现质的提升。

1.4　基于VLM的方法

VLM打破了传统的一对多分类任务模式，通过对比学习将视觉和文本模态的信息映射到统一的表征空间并对齐。其强大的多模态学习能力有效提升了模型对视频内容的理解能力，从而为视频动作识别任务带来显著的性能提升。CLIP[10]是一个通过4亿个图像文本对联合训练的VLM，凭借卓越的视觉和文本模态对齐能力，成为当前视觉任务转移知识的主流模型。文献[11]将CLIP的知识转移至视频动作识别领域，大幅提升了识别准确率。文献[11]的成功表明，基于VLM的视频动作识别方法具有巨大的探索空间，因此越来越多的优秀工作涌现而出。Ju等[32]提出一种在最少训练成本下高效使用预训练视觉—语言模型的视频动作识别基线方法。为了更有效地将VLM扩展到视频领域，Ni等[33]冻结了CLIP的参数，并提出一种跨帧注意力机制，用于捕捉帧时间维度上的长时间依赖关系。Pan等[34]提出一种时空适配器（ST-Adapter,

spatio-temporal adapter），在实现高效微调的同时提升了视频任务的时空推理能力。Zhao等[35]提出流视觉转换器（S-ViT, streaming vision transformer），将视频理解任务统一到一个新的流视频架构中，为视频动作识别框架的构建奠定了基础。文献[36]探索了线性分类器对于VLM将知识转移到视频领域的作用，尝试用预训练模型的不同知识替换分类器，从而实现高效的迁移学习。预训练VLM的核心价值在于构建视觉与文本间的桥梁，Wu等[37]提出一种框架，利用跨模态桥探索双向知识，结合视频属性关联机制，利用视频到文本的知识生成文本辅助属性进行视频识别。为保持模型的高监督性能和鲁棒可移植性，Wang等[38]在视觉和文本分支中引入多模态适配器，执行全局时间增强和局部时间差分建模，以提高视觉编码器的时间表示能力。这类模型通过结合动作类别的语义信息，突破了传统方法的性能瓶颈。然而，如何进一步提升对细粒度类别的识别能力仍是一个亟待解决的问题。

2　方法设计

为了提升模型对视频动作的细粒度区分能力，本文提出一种增强负例区分范式，并基于该范式设计了一种用于视频动作识别的多粒度区分模型（MGDM）。增强负例区分范式包括增强负例区分范式构建和增强负例集生成。MGDM框架如图2所示，该框架以增强负例区分范式为基础，主要包括视频特征提取模块和文本特征提取模块。本文所设计的视频表征器赋予视频特征提取模块关键的时序建模能力，而正负语义区分器则为文本特征提取模块提供了丰富的语义区分表征。本节主要从增强负例区分范式构建、增强负例集生成、视频表征器设计和正负语义区分器设计4部分介绍本文的主要方法。

2.1　增强负例区分范式构建

在视频动作识别任务样例中，包含一系列语义相似的场景或行为，这些与视频相关的类别容易导致模型对视频动作的误判。为提升模型对同一视频与其相关语义类别的区分能力，本文基于视频动作识别任务提出一种增强负例区分范式，如图1所示。与传统范式不同，本文在输入阶段为传统范式增加了最难区分的负例集，并将其扩展到每个视频-文本正例对内部。同时，增强真值标签也进行相应扩展。在传统范式的基础上，增强负例区分范式通过利用最难区分的部分负例集，提升了模型对同一视频与其相关语义类别的区分度。

在传统范式中，给定基础视频-文本对数据集 $\mathrm{DS}=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中，$(x_1,y_1)$为一个视频-文本正例对，$N$为数据集中总样本对数量。而增强负例范式的文本输入包括类别标签和增强负例集，则数据集重定义为

$$\widehat{\mathrm{DS}}=\{(x_1,\hat{y}_1),(x_2,\hat{y}_2),\cdots,(x_N,\hat{y}_N)\} \tag{1}$$

其中，$\hat{y}_1$为类别标签与增强负例集的集合，定义为

$$\hat{y}_1=[y_1,y_1^1,y_1^2,\cdots,y_1^k] \tag{2}$$

其中，$y_1^1\sim y_1^k$是长度为k的增强负例集，即对应视频x_1最难区分的部分文本负例。令X_{V}为视频输入，Y_{T}为文本输入，则模型优化目标函数为

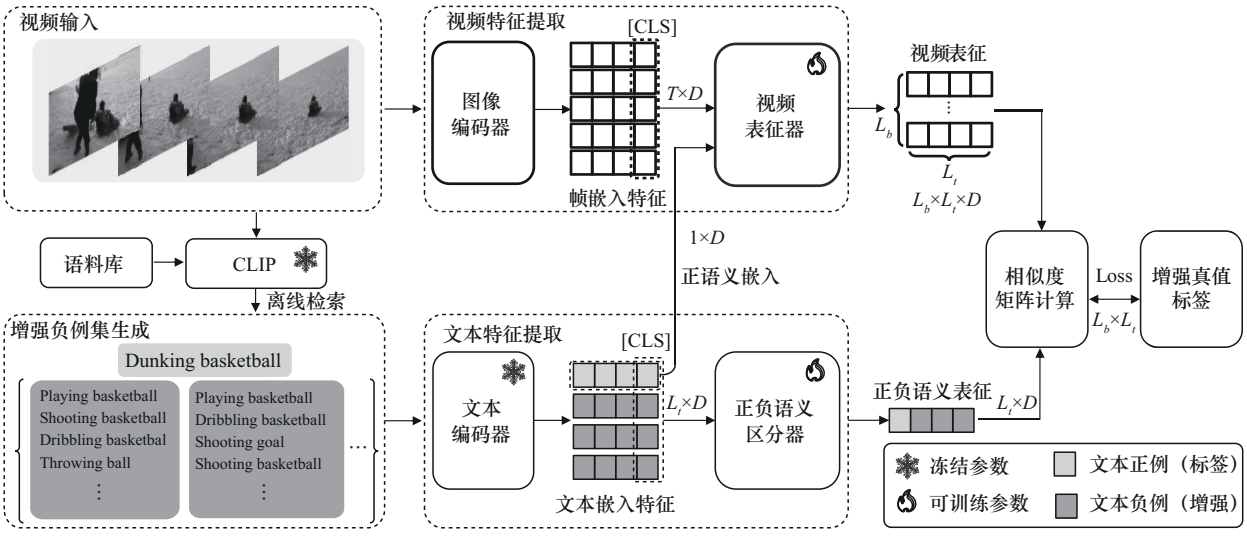


图2　基于增强负例区分范式的MGDM框架

$$\overset{*}{G}_{\mathrm{V}},\overset{*}{G}_{\mathrm{T}}=\underset{\Theta_{\mathrm{V}},\Theta_{\mathrm{T}}}{\arg\max}\,\mathbb{E}_{(X_{\mathrm{V}},Y_{\mathrm{T}})\sim\widehat{\mathrm{DS}}}[H(\boldsymbol{Q}|\sigma(G_{\mathrm{V}}(X_{\mathrm{V}})\otimes G_{\mathrm{T}}(Y_{\mathrm{T}})))] \tag{3}$$

其中，G_{V}和G_{T}分别为视频编码器和文本编码器，$\otimes$表示爱因斯坦求和，$\sigma(\cdot)$为softmax运算，$\boldsymbol{Q}$为增强后的真值标签，$H(\hat{P}|P)$为预测分布P和真实分布$\hat{P}$间的交叉熵，$\mathbb{E}$为期望，Θ_{V}和Θ_{T}分别为视频编码器和文本编码器的参数，$\overset{*}{G}_{\mathrm{V}}$和$\overset{*}{G}_{\mathrm{T}}$为最佳参数下的视频编码器和文本编码器。图1(b)中的相似度计算模块对应图2中的相似度矩阵计算。

令同一批次下视频和文本输入分别为$X_{\mathrm{V}}^{\mathrm{bs}}=[x_1,x_2,\cdots,x_b]$和$Y_{\mathrm{T}}^{\mathrm{bs}}=[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_l]$。在训练时$b=l$，即$b$为batch size；在推理时$b\neq l$，即$b$仍为batch size，而$l$为数据集中所有动作类别的数量。则$G_{\mathrm{V}}(X_{\mathrm{V}}^{\mathrm{bs}})\in\mathbb{R}^{L_b\times L_t\times D}$，$G_{\mathrm{T}}(Y_{\mathrm{T}}^{\mathrm{bs}})\in\mathbb{R}^{L_t\times D}$，其中，$L_b=b\times(1+k)$，$L_t=l\times(1+k)$。相似度矩阵计算方法为

$$\mathbf{sims}=G_{\mathrm{V}}(X_{\mathrm{V}}^{\mathrm{bs}})\otimes G_{\mathrm{T}}(Y_{\mathrm{T}}^{\mathrm{bs}})\in\mathbb{R}^{L_b\times L_t} \tag{4}$$

在训练阶段$b=l$，相似度矩阵**sims**与$\boldsymbol{Q}$为对应关系，**sims**的结构为

$$\mathbf{sims}=\begin{bmatrix} x_1y_1 & x_1y_1^1 & \cdots & x_1y_1^k & \cdots & x_1y_b & x_1y_b^1 & \cdots & x_1y_b^k(y_1)\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ x_1y_1 & x_1y_1^1 & \cdots & x_1y_1^k & \cdots & x_1y_b & x_1y_b^1 & \cdots & x_1y_b^k(y_1)\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ x_ly_1 & x_ly_1^1 & \cdots & x_ly_1^k & \cdots & x_ly_b & x_ly_b^1 & \cdots & x_ly_b^k(y_1)\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ x_ly_1 & x_ly_1^1 & \cdots & x_ly_1^k & \cdots & x_ly_b & x_ly_b^1 & \cdots & x_ly_b^k(y_1) \end{bmatrix} \tag{5}$$

其中，$\mathbf{sims}\in\mathbb{R}^{[b\times(1+k)]\times[b\times(1+k)]}$，$y_b^k(y_1)$表示$y_b^k$与$y_1$为同一类别标签，则与之对应的增强真值标签结构为

$$\boldsymbol{Q}=\begin{bmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & \cdots & 1\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots\\ 0 & 0 & \cdots & 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix} \tag{6}$$

定义一个批次中第i个视频-文本对为$(X_{\mathrm{V}i},Y_{\mathrm{T}i})$，则其在该批次中的概率分数为

$$P_{\mathrm{v2t}}(X_{\mathrm{V}i})=\frac{\exp\left(\frac{\mathbf{sims}[i][i]}{\tau}\right)}{\sum_{j=1}^{b}\exp\left(\frac{\mathbf{sims}[i][j]}{\tau}\right)} \tag{7}$$

$$P_{\mathrm{t2v}}(Y_{\mathrm{T}i})=\frac{\exp\left(\frac{\mathbf{sims}^{\mathrm{T}}[i][i]}{\tau}\right)}{\sum_{j=1}^{b}\exp\left(\frac{\mathbf{sims}^{\mathrm{T}}[i][j]}{\tau}\right)} \tag{8}$$

其中，τ为温度系数，$\mathbf{sims}[i][i]\in\mathbb{R}^{[1\times(1+k)]\times[1\times(1+k)]}$，$\mathbf{sims}^{\mathrm{T}}$是**sims**的转置，$P_{\mathrm{v2t}}(X_{\mathrm{V}i})$是视频到文本的相似度分数，$P_{\mathrm{t2v}}(Y_{\mathrm{T}i})$是文本到视频的相似度分数。本文结合式(6)定义的增强真值标签和KL散度，计算损失为

$$L=\frac{1}{2}\mathbb{E}_{(X_{\mathrm{V}},Y_{\mathrm{T}})\sim\widehat{\mathrm{DS}}}[\mathrm{KL}(P_{\mathrm{v2t}}(X_{\mathrm{V}}),Q_{\mathrm{v2t}}(X_{\mathrm{V}}))+\mathrm{KL}(P_{\mathrm{t2v}}(Y_{\mathrm{T}}),Q_{\mathrm{t2v}}(Y_{\mathrm{T}}))] \tag{9}$$

其中，$E_{(X_{\mathrm{V}},Y_{\mathrm{T}})\sim\widehat{\mathrm{DS}}}$表示数据集$\widehat{\mathrm{DS}}$中所有样例对的损失平均值，$L$为增强负例区分范式的损失，用于区分模态间视频和增强负例集的粗粒度。

在推理阶段$k=0$，b为batch size，l为数据集中所有动作类别的数量，即$\mathbf{sims}\in\mathbb{R}^{b\times l}$。考虑训练和推理2种情况下$b$和$l$的关系，因此本文在视频表征矩阵中保留$L_b$和$L_t$维度。

2.2　增强负例集生成

本文提出的增强负例区分范式的核心在于增强负例集生成，即为每个视频生成最难区分的负例集，这些负例集对应于图1(b)范式中文本编码器的输入。增强负例集生成过程主要包括视频输入、冻结参数的CLIP模型、语料库和离线检索4部分。如图2所示，通过采样一系列帧以代表该视频的信息，基于主流公开的视频动作识别数据集标签列表[13,39]构建语料库，将采样的视频帧与语料库同时输入冻结参数的CLIP模型。基于CLIP的强大视觉—语言特征对齐能力，MGPM模型可以为视频动作离线检索高置信度的语义先验区间。

经典的基于多模态的视频动作分类任务的概率

被建模为$P(s(x,y)|\theta)$，标签$s(\cdot)$是相似度计算函数。给定一组视频-文本对(x_1,y_1)，当x_1属于c_p类别时，y_1和c_p等同，该视频文本对也可表示为(x_1,c_p)。通过冻结CLIP的网络参数θ，视频x_1被离线检索为c_p类别的概率为

$$S_1^p = P(s(x_1,c_p)|\theta^*) \tag{10}$$

其中，S_1^p为第一个视频与其正确动作类别的分类概率，$*$表示网络参数θ为冻结状态。将语料库中的n个类别标签定义为$[c_p,c_1,c_2,\cdots,c_{n-1}]$，则有

$$S_1^i = P(s(x_1,c_i)|\theta^*) \tag{11}$$

其中，S_1^i为视频样例x_1与标签之外$n-1$个类别的分类概率，$i \in (1,2,\cdots,n-1)$。则在CLIP的先验知识下，x_1最难被区分的负例标签索引i为

$$i = \underset{1 \leqslant i \leqslant n-1}{\arg\max}\,(S_1^i) \tag{12}$$

依此类推，得出前k个最难区分的负例，从而组成该视频的增强负例集，即式(2)中的$[y_1^1,y_1^2,\cdots,y_1^k]$。其中，$y_1^1 \sim y_1^k$为$c_1 \sim c_{n-1}$中的不重复元素。

2.3　视频表征器设计

帧级图像输入经过图像编码器后得到帧级特征输出，未在时序上建模导致视频特征仍无法被有效表征。在增强负例集存在的情况下，本文设计的视频表征器引入文本正例特征引导帧间时序建模，为区分视频与增强负例集输出一体化的视频表征，视频表征器的结构如图3(a)所示。

给定视频-文本正例对，对于视频输入，采样的T个视频帧为$F=[F_1,F_2,\cdots,F_T]$，经过CLIP图像编码器后，分别取每帧的[CLS]特征得到帧嵌入特征$\boldsymbol{f}=[f_1,f_2,\cdots,f_T]$，其中$\boldsymbol{f} \in \mathbb{R}^{T\times D}$；对于文本输入，动作类别名称经过CLIP文本编码器后，取其[CLS]得到正语义嵌入$\boldsymbol{w}^p \in \mathbb{R}^{1\times D}$。

如图3(a)所示，帧嵌入与正语义嵌入作为视频表征器的输入。左支路输入帧嵌入特征$\boldsymbol{f}$经过M个Transformer模块时序建模后得到

$$\boldsymbol{f}^M = \text{block}_{\text{Trans}}(\boldsymbol{f}^{M-1}) \in \mathbb{R}^{T\times D} \tag{13}$$

其中，$\text{block}_{\text{Trans}}(\cdot)$表示一个Transformer模块，$\boldsymbol{f}^M$是$\boldsymbol{f}$经过$M$个Transformer模块的输出，然后进行残差连接后得到

$$\boldsymbol{f}^{\text{dM}} = \boldsymbol{f} + \boldsymbol{f}^M \tag{14}$$

其中，$\boldsymbol{f}^{\text{dM}} \in \mathbb{R}^{T\times D}$表示深度时序帧嵌入。将$\boldsymbol{f}^{\text{dM}}$张量扩展$L_b$份得到$\hat{\boldsymbol{f}}^{\text{dM}} \in \mathbb{R}^{L_b\times T\times D}$。

右支路输入正语义嵌入$\boldsymbol{w}^p$，将其张量扩展L_t份得到$\hat{\boldsymbol{w}}^p \in \mathbb{R}^{L_t\times D}$。对$\hat{\boldsymbol{f}}^{\text{dM}}$和$\hat{\boldsymbol{w}}^p$进行爱因斯坦求和约定计算，并映射到时序$T$上的注意力得到

$$\boldsymbol{V}_{\text{att}} = \text{softmax}((\hat{\boldsymbol{f}}^{\text{dM}} \otimes \hat{\boldsymbol{w}}^p), \text{dim}=-1) \in \mathbb{R}^{L_b\times L_t\times T} \tag{15}$$

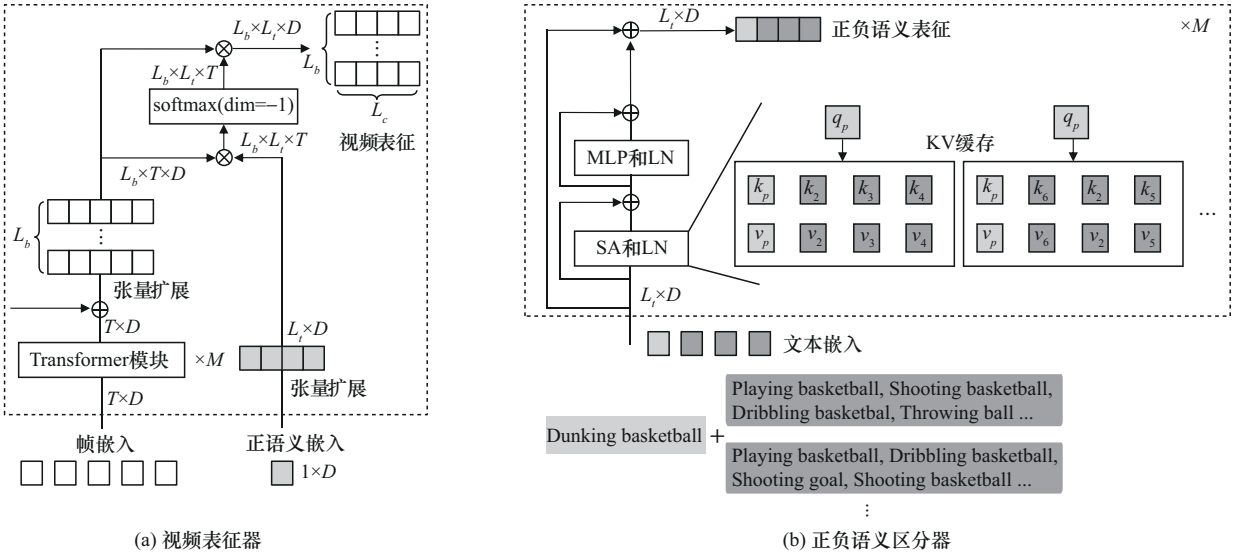
其中，$\otimes$表示爱因斯坦求和，$\text{dim}=-1$表示对T维度进行运算。根据时序注意力权重，计算视频表征为

$$\boldsymbol{V} = \hat{\boldsymbol{f}}^{\text{dM}} \otimes \boldsymbol{V}_{\text{att}} \in \mathbb{R}^{L_b\times L_t\times D} \tag{16}$$

其中，$\boldsymbol{V}$为一体化的视频表征矩阵。

2.4　正负语义区分器设计

从粗粒度的视频-文本表征角度考虑，利用正负语义区分度高的文本表征进行视频特征交互，以提升模型对同一视频与其相关语义类别的识别能



(a) 视频表征器　　(b) 正负语义区分器

图3　视频特征提取与文本特征提取模块

力；从细粒度的语义层面考虑，需要强化相关语义类别之间的区分度，以进一步提升文本正例与负例集之间的区分度。为此，本文设计了一个正负语义区分器，其结构如图3(b)所示。

同一视频动作类别对应大量视频样本，而同一类标签下不同视频样本对应不同的增强负例集合。对于同一类视频−文本对 (x_1,c_p) 和 (x_2,c_p)，通过2.2节检索出的增强负例集得到 $i=[p,2,3,4,\cdots,235]$ 和 $j=[p,6,2,5,\cdots,235]$。其中，i 和 j 代表同一动作类别下2个不同视频样本的负例索引集合，集合长度均为 $1+k$，p 表示正例标签索引，通过索引得到正例和负例集的文本，如图2中增强负例集所示。为每个类别构造一个语义句子（如“This is a video about dunking basketball”），并将该语义句子输入冻结参数的CLIP文本编码器，然后分别取每个类别句子的[CLS]特征，得到文本嵌入 $\boldsymbol{w}_i=[\boldsymbol{w}^p,\boldsymbol{w}^2,\boldsymbol{w}^3,\boldsymbol{w}^4,\cdots,\boldsymbol{w}^{235}]$，$\boldsymbol{w}_i\in\mathbb{R}^{L_t\times D}$，其中 $L_t=l\times(1+k)$，对于单个样例 $l=1$。

正负语义区分器输入的文本嵌入为 $\boldsymbol{w}$，当采样到视频−文本对 (x_1,c_p) 时，将 $\boldsymbol{w}_i$ 作为正负语义区分器的输入。在正负语义区分器中，Transformer模块的自注意力（SA, self-attention）机制能主动学习输入序列 $[\boldsymbol{w}^p,\boldsymbol{w}^2,\boldsymbol{w}^3,\boldsymbol{w}^4,\cdots,\boldsymbol{w}^{235}]$ 中文本特征正例 $\boldsymbol{w}^p$ 与增强负例集之间的自相关关系，从而实现文本模态内的细粒度区分。图3(b)展示了同类标签下2个样例的隐式正负语义关联机制示意图，通过正语义嵌入查询 q 与KV缓存区中的负例嵌入进行点积计算，得到概率分布为

$$\hat{\alpha}_{p,(i,j)}=\mathrm{softmax}\left(\frac{q_p k_{(i,j)}}{\sqrt{D}}\right) \tag{17}$$

其中，$i=[p,2,3,4,\cdots,235]$，$j=[p,6,2,5,\cdots,235]$，D 是嵌入向量的维度，$\hat{\alpha}_{p,(i,j,\cdots)}$ 为 q_p 与 $(i,j,\cdots)$ 负例集的自注意力概率分布，则自注意力加权输出为

$$e^p=\sum_i \hat{\alpha}_{p,(i,j)} * v_{(i,j)} \tag{18}$$

其中，$i=[p,2,3,4,\cdots,235]$、$j=[p,6,2,5,\cdots,235]$、e^p 为同一文本正例下 q_p 与不同视频样本对应的文本嵌入特征中所有key的自注意力输出。在式(17)和式(18)中，q、k、v 均来自 $\boldsymbol{w}$ 的特征映射。依此类推，输出的 $\boldsymbol{E}_i=[\boldsymbol{e}^p,\boldsymbol{e}^2,\boldsymbol{e}^3,\boldsymbol{e}^4,\cdots,\boldsymbol{e}^{235}]$ 和 $\boldsymbol{E}_j=[\boldsymbol{e}^p,\boldsymbol{e}^6,\boldsymbol{e}^2,\boldsymbol{e}^5,\cdots,\boldsymbol{e}^{235}]$ 分别为样例1和样例2经过一次SA计算后得到的文本表征，其中 $\boldsymbol{E}_i,\boldsymbol{E}_j\in\mathbb{R}^{L_t\times D}$，通过 M 个块级联后的输出与正负语义区分器的输入进行残差连接，得到最终的正负语义表征。

该正负语义区分器利用Transformer模块中的SA机制，形成 $\boldsymbol{w}^p$ 与 $\{[\boldsymbol{w}^2,\boldsymbol{w}^3,\boldsymbol{w}^4,\cdots,\boldsymbol{w}^{235}],[\boldsymbol{w}^6,\boldsymbol{w}^2,\boldsymbol{w}^5,\cdots,\boldsymbol{w}^{235}]\}$ 的自相关关系，主动建立 c_p 与增强负例集之间的区别与联系。对于 c_p 类别下的视频样本，该正负语义区分器可以建立文本模态内 c_p 与增强负例集并集之间更丰富的语义区分表征。

3　实验结果与分析

3.1　数据集

本文在3个广泛使用的动作识别数据集上对本文所提出的MGDM模型进行综合实验评估。

1) Kinetics-400数据集

Kinetics-400[12]是从YouTube上获取的数据集，该数据集包含400个人类日常生活活动类别，总计234 619个训练视频样本和19 761个验证视频样本，每个视频被裁剪到约10 s。本文实验采用Kinetics-400数据集的训练集进行模型训练，并在验证集上报告实验结果。

2) HMDB51数据集

HMDB51[39]是从电影和公共数据库等多个来源提取的数据集，该数据集包含51个行为类别，总计6 766个视频样本，每个类别至少包含101个片段，每个片段被裁剪到约5 s。文献[39]提供了3种将整个数据集拆分为训练集和测试集的方式。

3) UCF101数据集

UCF101[40]是从YouTube上获取的数据集，该数据集包含101个真实动作视频类，总计13 320个视频样本，每个视频被裁剪到约6 s。在101个动作类别中，视频被划分为25个组，同组视频可能具有一些共同特征，如相似背景或视角等。在UCF101数据集中，训练集包含9 537个样本，测试集包含3 783个样本。

3.2　实验设置与评价指标

1) 训练

在MGDM实验中，模型采用CLIP的视觉编码器[11]与时序编码器作为视频编码器，采用CLIP的文本编码器与正负语义编码器作为视频的文本编码器。对于视频输入，先从视频片段中采样 T 帧（$T=16$）图像，并将其短边大小调整为256，再采用随机缩放和裁剪以实现数据增强，将每帧调整为

224×224后输入网络。对于文本输入，先将正负语义类别标签构造成正负语义句子输入网络，再将正确类别标签特征用于视频的时序建模。此外，MGDM模型基于CLIP-400M预训练模型进行训练，并冻结或训练部分参数。在主要训练参数设置中，学习率为5×10^{-5}，优化器采用AdamW，模型在8×Tesla V100上进行训练，每个GPU的批次大小为32，预设定训练轮次为20，实验采用L2正则化方法缓解模型过拟合。

2) 评估

在MGDM实验中，对于视频输入，将视频帧短边大小调整为256，采用一次剪辑和中心裁剪模式，每次剪辑包含T帧（$T=16$）图像，每帧保留一个尺寸为224×224的中心裁剪。对于文本输入，将对应数据集的所有类别标签级联输入文本特征提取网络。通过计算验证集中视频样本与所有类别标签特征之间的相似度，进而计算得到对应数据集的Top-1和Top-5准确率。

3.3 负例集增强实验

为了对比图1中2种范式的性能，并突出增强负例集在视频动作识别任务中的有效性，分别为Kinetics-400、HMDB51和UCF101数据集设计负例集增强实验。根据范式中文本输入支路是否加入增强负例集，来测试不同负例数量对实验结果的影响，当无增强负例集时，负例个数k=0；当含负例集时，k=3,5,7,9，其中，k=3 (bot)表示选取Top-6与Top-9之间较易区分的负例，以体现不同难度负例对模型增强效果的影响。为更直观地展示含增强负例集与无增强负例集时识别精度的对比效果，括号中列出了含负例集与无负例集实验结果的差值。

在Kinetics-400数据集上的实验结果如表1所示，当无增强负例集时，Top-1准确率为85.9%，Top-5准确率为96.6%；而含增强负例集时，2类准确率先随着负例集元素数量的增加而增加，然后趋于平稳。当k=7时，含增强负例集相较于无增强负例集的Top-1准确率提升了1.4%，而Top-5准确率提升了1.3%，从而验证了增强负例区分范式的有效性。通过深入分析表1可得，仅使用视频和正例标签用于模型训练具有较好的Top-5性能，但其Top-1准确率较低，这表明传统基于视觉—语言范式的模型对细粒度动作区分的能力不足。通过为模型逐渐增加最难区分的一部分负例，模型在Kinetics-400数据集上的Top-1准确率提升显著，侧面表明增加的负例帮助模型提升了同一视频与其相关语义类别的区分能力。此外，通过对比k=3 (top)与k=3 (bot)可得，较难区分的负例在模型中的增强效果显著优于较易区分的负例。本文在对比实验中最终报告k=7的结果。

表1 Kinetics-400上的负例集增强实验

文本输入	k	Top-1	Top-5
无增强负例集	0	85.9%	96.6%
含增强负例集	3 (top)	86.6 (+0.7)%	97.3 (+0.7)%
	5 (top)	86.9 (+1.0)%	97.7 (+1.1)%
	7 (top)	87.3 (+1.4)%	97.9 (+1.3)%
	9 (top)	87.3 (+1.4)%	97.9 (+1.3)%
	3 (bot)	86.1 (+0.2)%	97.2 (+0.6)%

在HMDB51数据集上的实验结果如表2所示，当无增强负例集时，Top-1准确率为79.0%，Top-5准确率为96.8%；含增强负例集时，模型的Top-1准确率随着负例集元素数量的增加而增加，在k=7和k=9之间趋于平稳；而含增强负例集的Top-5准确率整体低于无增强负例集，主要原因在于HMDB51数据集中具有相似语义的类别较少，即类与类之间的距离较大。当视频采样帧中未提取到该动作的关键特征时，在推理阶段该视频会被高置信地归为错误类，而与之共同出现在增强负例集中的正确类的特征则会被推得更远，甚至超出Top-5的范围。与Kinetics-400数据集上的实验结果相同，为模型增加了最难区分的负例集后，其Top-1准确率显著提升，这进一步验证了增强负例集能有效提升模型对同一视频与其相关细粒度类别标签的识别能力。而k=3 (bot)的增强效果远低于k=3 (top)，这表明越难区分的负例对模型的增强效果越明显。本文在对比实验中最终报告k=7的结果。

表2 HMDB51上的负例集增强实验

文本输入	k	Top-1	Top-5
无增强负例集	0	79.0%	96.8%
含增强负例集	3 (top)	81.4 (+2.4)%	95.8 (−1.0)%
	5 (top)	81.7 (+2.7)%	96.0 (−0.8)%
	7 (top)	82.2 (+3.2)%	96.0 (−0.8)%
	9 (top)	82.0 (+3.0)%	96.1 (−0.7)%
	3 (bot)	79.7 (+0.7)%	96.5 (−0.3)%

在UCF101数据集上的负例集增强实验结果如表3所示，当无增强负例集时，Top-1准确率为96.3%，Top-5准确率为99.7%；含增强负例集时，随着增强负例数量的变化，模型的Top-5准确率未出现明显变化且都趋于1。当k=3 (bot)时，模型的Top-1准确率仅提升了0.6%，这表明越容易区分的负例对模型的增强效果越不明显。当k分别为7和9时，模型的Top-1准确率分别提升了1.8%和1.9%，且逐渐趋于平稳，这表明MGDM有效提取到了动作特征，且推开了正负例对距离。这也表明在UCF101数据集中最佳负例数为7~9，继续增加负例数量不会使视频动作识别的性能显著提升。虽然k=9时MGDM在UCF101数据集上的识别表现最佳，但为了统一参数，本文在对比实验中最终报告k=7的结果。

表3　　UCF101上的负例集增强实验

文本输入	k	Top-1	Top-5
无增强负例集	0	96.3%	99.7%
含增强负例集	3 (top)	97.5 (+1.2)%	99.7 (+0)%
	5 (top)	97.8 (+1.5)%	99.7 (+0)%
	7 (top)	98.1 (+1.8)%	99.7 (+0)%
	9 (top)	98.2 (+1.9)%	99.7 (+0)%
	3 (bot)	96.9 (+0.6)%	99.4 (−0.3)%

3.4　方法比较

为了验证MGDM的有效性，本文以预训练数据集作为对比方法的分类。其中，在Kinetics-400和HMDB51数据集上的对比方法不一致。

对于数据集Kinetics-400的实验，将MGDM与基于ImageNet[41]和基于CLIP-400M[11]的方法进行比较。基于ImageNet预训练的方法包括TSM[9]、STM[27]、TDN[30]、TAM[29]、TCM[42]、STM[43]、MT-Net[44]、AGPN[45]等，基于CLIP-400M预训练的方法包括ActionCLIP[10]、X-CLIP-B/16[33]、ST-Adapter[34]、VideoPrompt[32]、S-ViT-B/16[35]、Text4Vis[36]、BIKE[37]、M^2-CLIP-B/16[38]等。表4列出了不同方法的Top-1和Top-5准确率对比结果，对比方法的性能均来自对应文献。此外，视频的帧数与空间尺寸均会影响实验结果，为了保证对比的公平性，除了Text4Vis外，表4中的方法均统一了帧数与空间尺寸。

表4　　不同方法在Kinetics-400上的准确率

方法	预训练	输入	Top-1	Top-5
TSM[9]	ImageNet	16×224^2	74.1%	91.2%
STM[27]	ImageNet	16×224^2	73.7%	91.6%
TDN[30]	ImageNet	16×224^2	79.4%	94.4%
TAM[29]	ImageNet	16×224^2	79.3%	94.1%
TCM[42]	ImageNet	16×224^2	77.4%	93.1%
STM[43]	ImageNet	16×224^2	76.9%	92.7%
MT-Net[44]	ImageNet	16×224^2	78.1%	93.7%
AGPN[45]	ImageNet	16×224^2	77.7%	93.4%
ActionCLIP[10]	CLIP-400M	16×224^2	82.6%	96.2%
X-CLIP-B/16[33]	CLIP-400M	16×224^2	84.7%	96.8%
ST-Adapter[34]	CLIP-400M	16×224^2	86.9%	97.6%
VideoPrompt[32]	CLIP-400M	16×224^2	76.9%	93.5%
S-ViT-B/16[35]	CLIP-400M	16×224^2	84.7%	96.8%
Text4Vis[36]	CLIP-400M	32×224^2	87.1%	97.4%
BIKE[37]	CLIP-400M	16×224^2	87.2%	97.7%
M^2-CLIP-B/16[38]	CLIP-400M	16×224^2	83.7%	96.7%
MGDM	CLIP-400M	16×224^2	87.3%	97.9%

从表4可得，MGDM的Top-1和Top-5准确率均优于当前较具代表性的方法。整体来看，基于CLIP-400M预训练的方法均优于基于ImageNet预训练的方法，这表明预训练数据量的大小能显著影响模型的性能。在基于ImageNet预训练的方法中，表现最差的是STM[27]，尽管该方法对视频的时间−空间特征进行了建模，但由于网络深度不足，其性能受到限制；表现最好的是TDN[30]，该方法通过捕获多尺度时间信息进行动作识别，其Top-1和Top-5准确率分别达到79.4%和94.4%，在基于ImageNet预训练的方法中表现最佳。在基于CLIP-400M预训练的方法中，ActionCLIP[10]是早期具有代表性的方法之一，其在ViT-B/16版本下的Top-1准确率达到82.6%。表现最差的方法是VideoPrompt[32]，该方法通过为视频构造提示向量来弥补静态图像和视频之间的差距，但其过度依赖CLIP的零样本能力，没有重新建立视频与文本之间的联系；表现最好的方法是BIKE[37]，其利用CLIP建立视频−文本和文本−视频之间的双向桥梁，以增强视频表示，该方法的Top-1和Top-5准确率分别达到87.2%和97.7%，在基于CLIP-400M预训练的对比方法中表现最佳。MGDM的Top-1和Top-5

准确率分别为87.3%和97.9%，在基于ImageNet预训练的方法和基于CLIP-400M预训练的方法中均表现最佳。因此，在视频动作识别任务中，增强负例区分范式的有效性得到验证。

对于数据集HMDB51和UCF101的实验，本文将MGDM与基于Kinetics-400微调和基于ImageNet + Kinetics-400的方法进行比较。基于Kinetcs-400微调的方法包括DSDMT[46]，基于ImageNet + Kinetics-400的方法包括STM[27]、TSM[9]、TCM[41]、MT-Net[43]、STM[43]、CoViFocus[47]、STANet[48]、TTSN[49]等。表5和表6分别列出了2个数据集上不同方法的Top-1和Top-5准确率对比，对比方法的性能均来自对应文献。此外，视频的帧数与空间尺寸均会影响实验结果，为了保证对比的公平性，除了DSDMT和CoViFocus外，表5和表6中的方法均统一帧数与空间尺寸。

不同方法在HMDB51数据集上的实验结果如表5所示。由表5可知，基于Kinetics-400微调的方法仅有DSDMT[46]，原因在于该方法的输入模态包含帧、残差和运动矢量，无法直接在ImageNet预训练好的ResNet上进行微调，其报告的Top-1准确率为74.9%。在HMDB51上的主要对比方法类型为基于ImageNet + Kinetics-400微调的方法，其中表现最差的方法是STM[34]，与在Kinetcs-400数据集上的表现效果相似，由于网络深度不足，限制了该方法的性能；表现最好的对比方法为TTSN[49]，该方法提供了时序Transformer模块和时序自监控模块，利用有效的时间变换模块来建模非局部帧之间的非线性时间依赖性，显著增强了复杂运动特征的表示，其报告的Top-1准确率为80.2%。本文提出的MGDM模型报告的Top-1和Top-5准确率分别为82.2%和96.0%，在表5所

表5 不同方法在HMDB51上的准确率

方法	预训练	输入	Top-1	Top-5
DSDMT[46]	Kinetics-400	16×256^2	74.9%	—
STM[27]	ImageNet + Kinetics-400	16×224^2	72.2%	—
TSM[9]	ImageNet + Kinetics-400	16×224^2	73.5%	94.3%
TCM[42]	ImageNet +Kinetics-400	16×224^2	77.5%	—
MT-Net[44]	ImageNet + Kinetics-400	16×224^2	74.0%	—
STM[43]	ImageNet + Kinetics-400	16×224^2	75.2%	—
CoViFocus[47]	ImageNet + Kinetics-400	8×224^2	74.8%	—
STANet[48]	ImageNet + Kinetics-400	16×224^2	77.7%	—
TTSN[49]	ImageNet + Kinetics-400	16×224^2	80.2%	—
MGDM	CLIP-400M	16×224^2	82.2%	96.0%

表6 不同方法在UCF101上的准确率

方法	预训练	输入	Top-1	Top-5
DSDMT[46]	Kinetics-400	16×256^2	95.8%	—
STM[27]	ImageNet + Kinetics-400	16×224^2	96.2%	—
TSM[9]	ImageNet + Kinetics-400	16×224^2	95.9%	99.7%
TCM[42]	ImageNet + Kinetics-400	16×224^2	97.1%	—
MT-Net[44]	ImageNet + Kinetics-400	16×224^2	96.5%	—
STM[43]	ImageNet + Kinetics-400	16×224^2	97.1%	—
CoViFocus[47]	ImageNet + Kinetics-400	8×224^2	95.8%	—
STANet[48]	ImageNet + Kinetics-400	16×224^2	97.6%	—
TTSN[49]	ImageNet + Kinetics-400	16×224^2	96.8%	—
MGDM	CLIP-400M	16×224^2	98.1%	99.7%

有实验方法中表现最好，这表明MGDM能更有效地对齐视频-文本对。

表6展示了不同方法在UCF101数据集上的准确率。DSDMT[46]和CoViFocus的Top-1准确率均为95.8%，是表6中Top-1准确率最低的。DSDMT的预训练数据集中不包含ImageNet，导致预训练知识不足；CoViFocus只采样了8帧，这可能导致错过关键视频帧。表现最好的对比方法为STANet[48]，该方法通过时间自适应模块和空间自适应模块，对关键时序帧和关键空间信息进行建模，很好地平衡了动作的静态语义和动态运动，其报告的Top-1准确率达到97.6%。本文提出的MGDM的Top-1和Top-5准确率分别为98.1%和99.7%，均为所有对比方法中最高的，这表明MGDM能够有效提取视频和文本特征，并强化视频-文本对之间的特征对齐能力。

3.5　消融研究

3.3节验证了增强负例区分范式及负例数对模型结果的影响，在此基础上，为分析MGDM中的模块对最终结果的影响，本节为视频表征器和正负语义区分器设计了消融实验。表7展示了在3个数据集上，MGDM分别移除视频表征器和正负语义区分器的实验结果。视频表征器的作用是将帧嵌入特征转换为视频特征，移除该表征器后会显著影响视频特征提取，因此在3个数据集上无视频表征器时，动作识别Top-1准确率均明显低于MGDM。正负语义区分器的作用是主动构建正负语义类别之间的自相关关系，以关注文本模态正例与增强负例之间的细粒度区分，移除该模块会影响语义层面的类别区分，但不会影响视觉和文本的特征提取，因此在3个数据集上无正负语义区分器时，动作识别Top-1准确率均小幅低于MGDM。总体而言，缺失视频表征器会影响视频特征提取，而缺失正负语义区分器会影响文本模态内正负例间的细粒度区分，2个模块对于MGDM都是不可或缺的。

表7　Kinetics-400、HMDB51、UCF101上的Top-1准确率

方法	Kinetics-400	HMDB51	UCF101
无视频表征器	83.7 (−3.6) %	76.6 (−5.6) %	96.5 (−1.6) %
无正负语义区分器	86.2 (−1.1) %	80.1 (−2.1) %	97.3 (−0.8) %
MGDM	87.3%	82.2%	98.1%

为了评估MGDM的复杂度，本节开展了每秒浮点运算次数（FLOP）和参数量对比，实验结果如表8所示。骨干网络CLIP ViT-L/14的FLOP和参数量分别为834.7×10^9和258.7×10^6，在此基础上设计了MGDM。当k=0时，FLOP只增加了0.4×10^9，参数量只增加了37.8×10^6；当k=7时，FLOP相比CLIP和k=0时分别增加了0.5×10^9和0.1×10^9，参数量与k=0时一致。通过上述分析可得，MGDM相较于CLIP所增加的FLOP和参数量较低，且当增强负例集数量增加时，仅FLOP有细微增加，计算成本较低。

表8　模型复杂度分析

方法	输入	FLOP	参数
CLIP ViT-L/14	16×224^2	834.7×10^9	258.7×10^6
MGDM (k=0)	16×224^2	$835.1\ (+0.4)\times10^9$	$296.5\ (+37.8)\times10^6$
MGDM (k=7)	16×224^2	$835.2\ (+0.5)\times10^9$	$296.5\ (+37.8)\times10^6$

3.6　可视化分析

MGDM模型虽然实现了视频-文本的模态对齐，但其特征仍无法映射到同一空间，因此无法直接采用通用的特征降维算法。为了更好地展示视频-文本在特征层面的对齐效果，同时更直观地看出文本模态内语义相关类别特征之间的区分度增强效果，本节设计的可视化方式如图4所示。D为向量的长度，sim为相似度计算，通过将D维特征向量切分为长度相等的2段，并从左右2端分别计算相似度，左端的相似度定义为横坐标，右端的相似度定义为纵坐标。该可视化方式可以将维度为D的特征投射到2维坐标空间，同时可体现视频-文本对齐的程度，坐标越靠近右上角，表明该语义特征与视频特征越匹配。倒三角坐标与圆点距离越远，表明正语义特征与负语义特征的区分度越高。

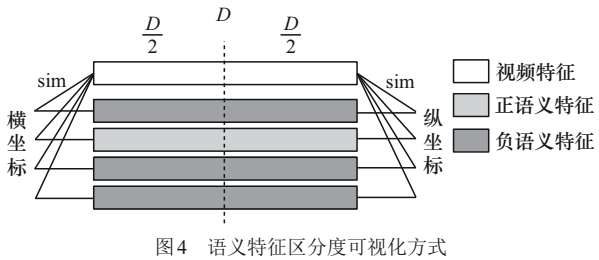


图4　语义特征区分度可视化方式

图5显示了在Kinetics-400、HMDB51和UCF101数据集上语义特征区分度的可视化结果，其中视频表征、文本嵌入特征和正负语义区分表征

(a1) Stretching Leg (a2) Stretching Leg 表征与文本嵌入特征 (a3) Stretching Leg 表征与正负语义表征

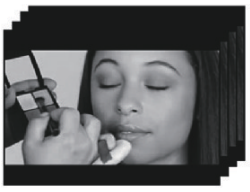
(a4) Somersaulting (a5) Somersaulting 表征与文本嵌入特征 (a6) Somersaulting 表征与正负语义表征

(a) Kinetics-400数据集可视化样例

(b1) Brush Hair (b2) Brush Hair表征与文本嵌入特征 (b3) Brush Hair表征与正负语义表征

(b4) Hug (b5) Hug表征与文本嵌入特征 (b6) Hug表征与正负语义表征

(b) HMDB51数据集可视化样例



(c1) Apply Eye Makeup (c2) Apply Eye Makeup表征与文本嵌入特征 (c3) Apply Eye Makeup表征与正负语义表征

(c4) HulaHoop (c5) HulaHoop表征与文本嵌入特征 (c6) HulaHoop表征与正负语义表征

(c) UCF101数据集可视化样例

图5 在Kinetics-400、HMDB51和UCF101数据集上的可视化结果

来自图2，倒三角坐标表示视频与标签的对齐程度，圆点坐标表示视频与非标签的对齐程度。从图5中可以得出，在Kinetics数据集展示的样例中，仅使用视频表征与文本嵌入特征使得倒三角坐标混淆进入圆点坐标中，即模型无法将视频与其标签对应。而在视频表征与语义区分表征列中，在保持视频表征不变的情况下，倒三角坐标明显远离圆点聚集的区域。在HMDB51数据集的可视化样例中，文本特征经过正负语义区分器后，倒三角坐标更显著地远离圆点聚集的坐标。在UCF101数据集中，圆点比较聚集，表明文本嵌入表征的区分度不高，而经过正负语义区分器后，文本模态中的类特征区分度更显著，并且推开了倒三角坐标与圆点坐标之间的距离。以上现象都体现了MGDM不仅能增强视频-文本的对齐程度，还能提升语义层面正例和负例的区分度。

4　结束语

本文提出一种增强负例区分范式，以增强视频与其相关语义类别的区分度，从而提升模型对细粒度动作的识别能力。基于增强负例区分范式，设计了多粒度区分模型MGDM，该模型通过引入文本正例特征引导帧间时序建模的视频表征器，以及通过自注意力机制主动构建正负语义自相关关系的正负语义区分器，同时实现了模态间和模态内2种粒度的区分。负例集增强实验结果表明，为模型增加最难区分的负例能有效提升视频动作识别的准确率。实验结果表明，多粒度区分模型在Kinetics-400、HMDB51和UCF101数据集上均表现出较高的Top-1和Top-5准确率，消融实验进一步验证了增强负例集和增强负例区分范式能有效提高模型对细粒度动作的识别能力。

尽管增强负例集能有效提升同一视频与其相关语义类别的区分度，但基于VLM的视频动作识别方法仍然面临对CLIP零样本能力的过度依赖问题。在分类CLIP预训练知识之外的样本时，现有方法表现不好。因此，如何为待识别的视频提供更多知识是未来的研究方向之一。本文后续研究将聚焦于多模态大模型在视频动作识别领域的应用，特别是针对CLIP在下游视觉任务场景中理解能力的局限性，探索更全面、更高效的解决方案。

参考文献：

[1] 顾晓丹, 吴文甲, 凌振. 用户密集环境下基于边缘智能的直播视频传输优化机制[J]. 通信学报, 2023, 44(11): 55-66.
GU X D, WU W J, LING Z. Live video transmission optimization mechanism based on edge intelligence in high client-density environment[J]. Journal on Communications, 2023, 44(11): 55-66.

[2] 毕春艳, 刘越. 基于深度学习的视频人体动作识别综述[J]. 图学学报, 2023, 44(4): 625-639.
BI C Y, LIU Y. A survey of video human action recognition based on deep learning[J]. Journal of Graphics, 2023, 44(4): 625-639.

[3] 张伟, 王宇, 陈新怡, 等. 基于图像块码本模型的监控视频背景参考帧生成方法[J]. 通信学报, 2023, 44(1): 129-141.
ZHANG W, WANG Y, CHEN X Y, et al. Background reference frame generation method for surveillance video based on image block codebook model[J]. Journal on Communications, 2023, 44(1): 129-141.

[4] 许可, 李嘉怡, 蒋兴浩, 等. 一种基于轮廓稀疏对抗的视频步态隐私保护算法[J]. 信息网络安全, 2024, 24(1): 48-59.
XU K, LI J Y, JIANG X H, et al. A video gait privacy protection algorithm based on sparse adversarial attack on silhouette[J]. Netinfo Security, 2024, 24(1): 48-59.

[5] NAM B T, JUNGER D, CURIO C, et al. Towards human action recognition during surgeries using de-identified video data[J]. Current Directions in Biomedical Engineering, 2022, 8(1): 109-112.

[6] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1725-1732.

[7] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[J]. Neural Information Processing Systems, 2014(1): 568-576.

[8] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 4724-4733.

[9] LIN J, GAN C, HAN S. TSM: temporal shift module for efficient video understanding[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 7082-7092.

[10] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. New York: ACM Press, 2021: 8748-8763.

[11] WANG M M, XING J Z, LIU Y. ActionCLIP: a new paradigm for video action recognition[J]. arXiv Preprint, arXiv: 2109.08472, 2021.

[12] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[J]. arXiv Preprint, arXiv: 1705.06950, 2017.

[13] CAI T T, FRANKLE J, SCHWAB D J, et al. Are all negatives created equal in contrastive instance discrimination?[J]. arXiv Preprint, arXiv: 2010.06682, 2020.

[14] HORN B K P, SCHUNCK B G. Determining optical flow[J]. Artificial Intelligence, 1981, 17(1-3): 185-203.

[15] WANG L M, QIAO Y, TANG X O. Action recognition with trajectory-

pooled deep-convolutional descriptors[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 4305-4314.

[16] DONAHUE J, HENDRICKS L A, GUADARRAMA S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2015: 2625-2634.

[17] WANG L M, XIONG Y J, WANG Z, et al. Temporal segment networks: towards good practices for deep action recognition[C]//Lecture Notes in Computer Science. Berlin: Springer, 2016: 20-36.

[18] FEICHTENHOFER C, PINZ A, WILDES R P. Spatiotemporal multiplier networks for video action recognition[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 7445-7454.

[19] JI S W, YANG M, YU K. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

[20] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 4489-4497.

[21] HARA K, KATAOKA H, SATOH Y. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? [C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6546-6555.

[22] XIE S N, SUN C, HUANG J, et al. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification[C]//Lecture Notes in Computer Science. Berlin: Springer, 2018: 318-335.

[23] FEICHTENHOFER C, FAN H Q, MALIK J, et al. SlowFast networks for video recognition[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 6201-6210.

[24] FEICHTENHOFER C. X3D: expanding architectures for efficient video recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 200-210.

[25] ZHU Y, LI X Y, LIU C H, et al. A comprehensive study of deep video action recognition[J]. arXiv Preprint, arXiv: 2012.06567, 2020.

[26] PIERGIOVANNI A, ANGELOVA A, RYOO M S. Tiny video networks[J]. arXiv Preprint, arXiv: 1910.06961, 2019.

[27] JIANG B Y, WANG M M, GAN W H, et al. STM: Spatiotemporal and motion encoding for action recognition[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2019: 2000-2009.

[28] LI Y, JI B, SHI X T, et al. TEA: temporal excitation and aggregation for action recognition[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 906-915.

[29] LIU Z Y, WANG L M, WU W, et al. TAM: temporal adaptive module for video recognition[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2021: 13688-13698.

[30] WANG L M, TONG Z, JI B, et al. TDN: temporal difference networks for efficient action recognition[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 1895-1904.

[31] ZHANG Y T, BAI Y, WANG H, et al. Look more but care less in video recognition[J]. Neural Information Processing Systems, 2022, 35: 30813-30825.

[32] JU C, HAN T D, ZHENG K H, et al. Prompting visual-language models for efficient video understanding[C]//Lecture Notes in Computer Science. Berlin: Springer, 2022: 105-124.

[33] NI B L, PENG H W, CHEN M H, et al. Expanding language-image pretrained models for general video recognition[C]//Lecture Notes in Computer Science. Berlin: Springer, 2022: 1-18.

[34] PAN J T, LIN Z Y, ZHU X T, et al. ST-adapter: parameter-efficient image-to-video transfer learning[J]. Neural Information Processing Systems, 2022, 35: 26462-26477.

[35] ZHAO Y C, LUO C, TANG C X, et al. Streaming video model[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 14602-14612.

[36] WU W H, SUN Z, OUYANG W L. Revisiting classifier: transferring vision-language models for video recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2023, 37(3): 2847-2855.

[37] WU W H, WANG X H, LUO H P, et al. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 6620-6630.

[38] WANG M M, XING J Z, JIANG B Y, et al. A multimodal, multi-task adapting framework for video action recognition[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(6): 5517-5525.

[39] KUEHNE H, JHUANG H, GARROTE E, et al. HMDB: a large video database for human motion recognition[C]//Proceedings of the 2011 International Conference on Computer Vision. Piscataway: IEEE Press, 2011: 2556-2563.

[40] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild[J]. arXiv Preprint, arXiv: 1212.0402, 2012.

[41] DENG J, DONG W, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2009: 248-255.

[42] LIU Y Z, YUAN J S, TU Z G. Motion-driven visual tempo learning for video-based action recognition[J]. IEEE Transactions on Image Processing, 2022, 31: 4104-4116.

[43] WANG M M, XING J Z, SU J, et al. Learning spatiotemporal and motion features in a unified 2D network for action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45 (3): 3347-3362.

[44] SHENG X X, LI K C, SHEN Z Q, et al. A progressive difference method for capturing visual tempos on action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33 (3): 977-987.

[45] CHEN Y T, GE H W, LIU Y X, et al. AGPN: action granularity pyramid network for video action recognition[J]. IEEE Transactions on Cir-

cuits and Systems for Video Technology, 2023, 33(8): 3912-3923.
[46] MOU Y T, JIANG X H, XU K, et al. Compressed video action recognition with dual-stream and dual-modal transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(5): 3299-3312.
[47] ZHENG Z W, YANG L, WANG Y L, et al. Dynamic spatial focus for efficient compressed video action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(2): 695-708.
[48] LI Z L, LI J, MA Y Q, et al. Spatio-temporal adaptive network with bi-directional temporal difference for action recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(9): 5174-5185.
[49] ZHANG Y K, LI J, JIANG N, et al. Temporal transformer networks with self-supervision for action recognition[J]. IEEE Internet of Things Journal, 2023, 10(14): 12999-13011.

[作者简介]

刘良振（1998– ），男，湖南邵阳人，中南大学博士生，主要研究方向为视频动作识别、视频行为监管。

杨阳（1999– ），男，安徽合肥人，中南大学博士生，主要研究方向为代码生成、智能运维、多模态学习。

夏莹杰（1982– ），男，浙江奉化人，博士，杭州电子科技大学特聘教授、浙江大学兼聘教授，主要研究方向为智能交通和信息安全。

邝砾（1982– ），女，湖南长沙人，博士，中南大学教授、博士生导师，主要研究方向为智能软件工程与服务监管。