基于对比学习和未来特征预测的早期行为识别方法

张洪博[1),3)] 陈 婕[1),2)] 郑博圣[3)] 刘景华[1)] 杜吉祥[2)] 孙真真[1)]

[1)](华侨大学计算机科学与技术学院 福建 厦门 361000)

[2)](华侨大学福建省大数据智能与安全重点实验室 福建 厦门 361000)

[3)](华侨大学厦门市计算机视觉与模式识别重点实验室 福建 厦门 361000)

摘 要 针对早期行为识别任务中因仅能观测到视频的初始片段、时序信息不完整而导致识别精度受限的问题，本文提出了一种融合对比学习和未来特征预测的早期行为识别方法。该方法基于Transformer架构，设计了面向未来不可观测片段的特征预测模型，并引入对比学习机制以提升预测特征的判别能力。具体而言，该模型首先利用视频特征编码器对可观测片段进行时序建模；随后，采用基于交叉注意力机制的解码器生成未来片段的特征，从而显式建模早期可观测片段与未来片段之间的映射关系。其次，在训练阶段，通过构建“预测未来特征—真实未来特征”作为正样本对、“预测未来特征—早期片段特征”作为负样本对的策略，引入对比损失函数，以增强预测特征与早期特征之间的区分度，进而提升模型性能。本文在Something-Something V2与UCF101两个公开数据集上进行了定量与定性分析。实验结果表明，所提方法在不同观测比例下均优于现有方法，在具有挑战性的Something-Something V2数据集上平均准确率提升了4.96%；在已有方法精度较高的UCF101数据集上，准确率进一步从89.44%提升至90.03%，充分验证了本文引入的对比学习策略与整体模型设计的有效性。同时，实验还表明该方法在提升性能的同时兼顾了模型参数规模，体现出良好的效率和性能平衡。

关键词 早期行为识别；对比学习；早期可观测片段；未来特征预测；Transformer模型

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.00447

Early Action Recognition Method Based on Contrastive Learning and Future Feature Prediction

ZHANG Hong-Bo[1),3)] CHEN Jie[1),2)] ZHENG Bo-Sheng[3)] LIU Jing-Hua[1)]
DU Ji-Xiang[2)] SUN Zhen-Zhen[1)]

[1)](Department of Computer Science and Technology, Huaqiao University, Xiamen, Fujian 361000)

[2)](Fujian Key Laboratory of Big Data Intelligence and Security, Huaqiao University, Xiamen, Fujian 361000)

[3)](Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen, Fujian 361000)

Abstract Early action recognition aims to classify ongoing activities using only initial video segments, a constraint that often leads to suboptimal accuracy due to incomplete temporal information. To address this, this paper proposes a novel method that integrates contrastive learning with future feature prediction. The proposed method is built on a Transformer-based encoder-decoder architecture, which includes a module dedicated to predicting features of

收稿日期：2025-04-28；在线发布日期：2025-11-02。本课题得到国家自然科学基金面上项目(No. 61871196)、国家自然科学基金青年科学基金C类项目(No. 62306121)、福建省自然科学基金面上项目(No. 2025J01177)、厦门市自然科学基金面上项目(No. 3502Z202373040)资助。 **张洪博**，博士，教授，中国计算机学会(CCF)高级会员，主要研究领域为计算机视觉、行为理解。E-mail：zhanghongbo@hqu.edu.cn。**陈 婕**，硕士研究生，主要研究领域为计算机视觉、早期行为识别。**郑博圣**，硕士研究生，主要研究领域为计算机视觉、早期行为识别。**刘景华**，博士，副教授，中国计算机学会(CCF)高级会员，主要研究领域为机器学习、数据挖掘。**杜吉祥**(通信作者)，博士，教授，中国计算机学会(CCF)会员，主要研究领域为人工智能、图像处理。E-mail：jxdu@hqu.edu.cn。**孙真真**，博士，讲师，中国计算机学会(CCF)会员，主要研究领域为机器学习、特征选择。

unobserved future segments. A key contribution is the introduction of a contrastive learning mechanism designed to enhance the discriminability of these predicted features. Specifically, the model first employs a video feature encoder to perform temporal modeling on the observable segments. Subsequently, a decoder based on a cross-attention mechanism is utilized to generate features for future segments, thereby explicitly modeling the mapping relationship between early observable segments and future ones. During training, a strategy is adopted that constructs positive sample pairs as "predicted future features-real future features" and negative sample pairs as "predicted future features-early segment features." A contrastive loss function is introduced to strengthen the distinction between predicted features and early features, thus improving overall model performance. Extensive quantitative and qualitative experiments are conducted on two public datasets, Something-Something V2 and UCF101. The results demonstrate that the proposed method outperforms existing approaches across various observation ratios. On the challenging Something-Something V2 dataset, it achieves an average accuracy improvement of 4.96%. On the UCF101 dataset, where baseline accuracy is already high, the proposed method further increases accuracy from 89.44% to 90.03%. These results adequately validate the effectiveness of the introduced contrastive learning strategy and the overall model design. Furthermore, experiments show that the method achieves this performance enhancement while maintaining a manageable parameter scale, reflecting a favorable balance between efficiency and performance.

Keywords early action recognition; contrastive learning; early observed segments; future feature prediction; Transformer model

1 引 言

早期行为识别(Early Action Recognition,EAR)是指在仅观察到视频起始部分的情况下,准确地识别出视频中即将发生的行为类别。与传统的完整视频行为识别任务相比,EAR任务的重点是对行为的提前预测,是视频分析领域中的一个重要研究方向。EAR不仅在异常事件预警和人机交互理解等应用中具有重要意义,而且还是实现多种实时智能监测系统的核心技术之一[1]。

在EAR任务中,首先将一个完整的视频分为两部分:早期片段和后期片段。其中,早期片段的长度比例通常被定义为"观察率",如图1所示。在模型的训练阶段,早期片段和后期片段都是已知的,并且都作为模型训练的输入。然而,在推理阶段,仅早期片段是可用的,后期片段未知。在这种情况下,模型需要基于早期片段的特征信息准确识别视频中的行为类别。由于早期片段中所包含的行为信息是片段性的且不完整的,缺乏后续行为信息,因此EAR任务相较于传统的行为识别任务更具挑战性。

面对视频数据不完整场景下的EAR任务,一种常见的解决方法是通过构建后期片段信息预测模型来弥补缺失的后期行为特征。具体而言,这些方法通过输入早期片段的视频特征,预测未知的后期片段特征,填补未来行为信息的空缺,进而提高早期行为识别的准确性。例如,Li等人[2]利用后期片段特征与早期片段特征的距离损失,提出了一种基于交叉注意力机制的后期特征补充网络。Oord等人[3]提出了一种基于对比预测编码(Contrastive Predictive Coding, CPC)的自监督学习方法,在数据稀缺的情况下有效为下游动作识别任务提供有意义的特征表示。Zhang等人[4]提出了一种对抗注意力网络,采用对抗学习机制训练从早期片段到后期片段的特征映射,并将预测的特征融合到行为识别网络中,从而提升了EAR模型的性能。

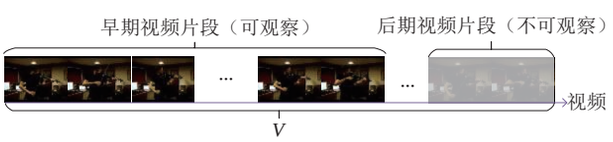

图1 早期视频片段和后期视频片段示意图

这些方法表明,未来特征预测可以有效提升EAR任务的识别精度。在这些方法的训练中,将后

期视频片段的特征作为标签，利用预测的特征与标签之间的相似度来约束未来特征生成模块的训练，从而弥补缺失的行为信息，如图2所示。行为通常由多个不同的子动作随时间演变构成，在不同时间节点上可能存在显著差异。另一方面，预测不可观察片段特征的主要目的在于补充早期片段中缺失的信息。如果预测得到的特征与早期特征高度一致，则说明其提供的增量信息有限，无法有效增强后续行为判别能力。因此，从行为演化特征和信息增益角度出发，期望预测的不可观察特征应与早期可观察特征存在一定差异。但是，目前多数工作仅关注如何拉近预测的未来特征与真实标签特征之间的距离，从而提升预测一致性，而往往忽略了预测未来特征与早期已观察特征之间必要的区分度，难以有效补充早期片段中缺失的重要时序信息。如何使得预测的未来特征更加具有差异化和判别性，进而提高EAR任务的准确性，仍是一个需要深入研究的问题。

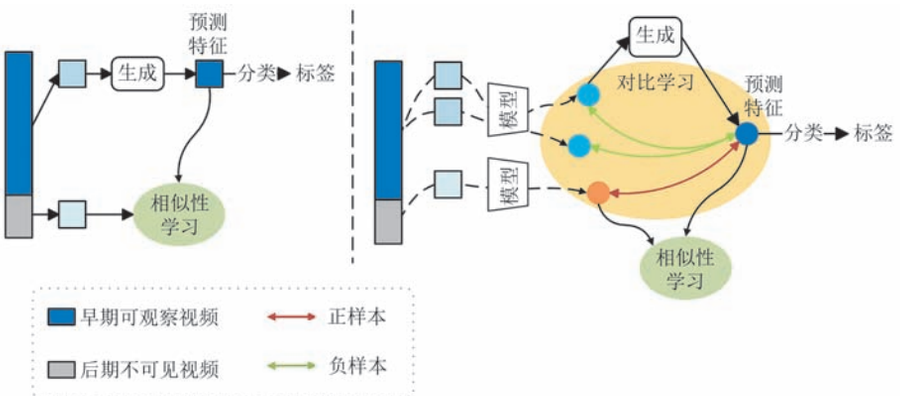

图2　传统特征预测(左)和对比学习的预测(右)

为了解决这个问题，提升预测特征的判别能力并增强其与早期特征的差异性，本文提出了一种基于对比学习的未来特征预测方法，其流程如图2所示。首先，本文通过构建行为分类网络实现预测特征与行为标签之间的映射，补充预测特征的行为语义，从而增强其判别性。其次，本文基于未来特征和早期特征构造正负样本对，提出了基于对比损失的未来特征学习方法，使模型聚焦于具有区别性的关键特征，进一步提高预测的未来特征的区分性。

基于此，本文设计了一种融合对比学习与Transformer架构的EAR模型，如图3所示。首先通过Swin Transformer从视频中提取特征序列；随后，将任务标识向量(token)作为可学习参数，通过通道拼接的方式与骨干网络提取的特征序列结合，并输入编码器模块，以捕捉已观察片段的全局时间依赖关系；解码器则负责预测未来可能发生的行为特征。最后，在解码阶段，将编码器输出的token与解码器预测的未来特征进行串联，形成完整的视频特征序列，用于后续的分类判别。结合对比学习和Transformer架构的优势，本方法进一步提升了EAR模型的精度。

尽管对比学习已广泛应用于图像与视频领域，如SimCLR[5]、MoCo[6]，感知推理对比方法[7]，以及时空对比视频表征学习方法[8]，但这些方法主要聚焦于自监督表示学习或完整视频的特征建模，较少涉及EAR任务中预测未来片段的问题。具体来说，现有对比学习工作侧重于样本之间的一致和差异性的特征学习，而在EAR场景下的未来特征建模不仅需关注预测未来特征与标签之间的一致性，同时还需体现由行为时序演变带来的差异性。因此，如何将对比学习机制有效地嵌入EAR任务中的特征生成与预测过程，仍是尚未被充分研究的问题。为此，本文设计了“预测片段-真实后期片段”为正样本对，“预测片段-早期可观察片段”为负样本对的对比学习方式。这种方式融合预测监督与时序区分性的对比策略，契合了EAR任务“信息不完整”与“判别性建模并重”的独特需求，有效增强预测特征的判别性，从而提升EAR任务的精度。

综上所述，本文的主要贡献如下：

(1)本文提出了一种基于对比学习的未来特征预测方法，将对比学习引入有监督的表征学习中。在该方法中，首先将可观测的早期片段划分为多个负样本，同时利用后期视频生成正样本。通过构建正负样本对的对比学习框架，模型旨在缩小正样本对间的差距，放大负样本对间的差异，从而使得预测

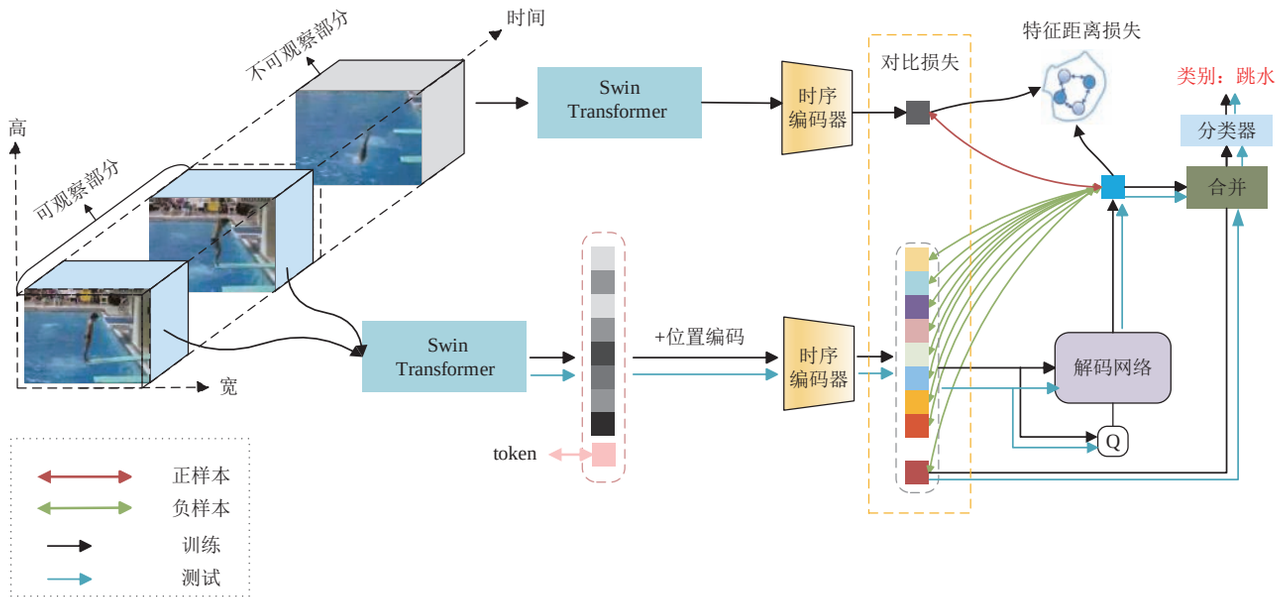

图3 本文所提出的模型框架图

特征更加接近真实的后期特征，同时增强与早期片段特征的差距。

(2)本文构建了一种融合对比学习和Transformer的EAR方法。该方法采用了编码器-解码器结构，使得模型能够在编码过程中有效地捕捉已知视频片段的时序特征，并在解码阶段通过融合对比学习和特征距离损失重建后期行为特征。最终，预测的未来特征与早期片段特征相融合，从而提升了EAR的精度。

(3)本文在公共数据集UCF101[9]和Something-Something V2[10]上进行了实验验证和分析。结果表明本文方法在EAR任务中表现出较高的准确性。此外，详细的消融实验进一步验证了本文所提出的对比学习策略能进一步提升EAR任务的精度。

2 相关工作

2.1 视频中的行为特征学习方法

视频行为特征学习和识别是计算机视觉领域的经典研究任务，传统方法主要包括基于三维时空卷积的特征编码[11]以及基于帧序列的时序编码[12]。随着Transformer模型在图像任务的成功应用，基于Transformer的行为特征学习方法成为研究热点。例如，ViViT算法[13]是将图像分类中的Vision Transformer模型的分类token拓展到了三维，称为时空令牌。该令牌被添加在由输入数据所组成的线性序列当中，跟随输入序列经过一系列Transformer层的学习后被用于最后的行为分类。Qing等[14]将视频掩码自编码模型应用到行为特征学习中，提出了掩码行为识别模型。Mou等[15]针对视频压缩域的数据特点，提出了双流和双模态Transformer模型。

除了对Transformer模型架构进行探索和优化外，也有学者从数据增强，模型训练以及模型压缩等方面对行为特征学习方法进行了研究。考虑到应用于视频数据的视觉模型大多遵循“图像预训练然后微调”范式，Yang等[16]提出了空间适应、时间适应和联合适应方式三种方式来自适应预训练的图像模型以实现高效的视频理解。为了增强视频帧之间的时序关系，消除不同模态特征之间的信息差异，吴沛宸等[17]提出基于特征增强和模态交互的视频异常行为检测算法。针对测试数据分布与训练数据不一致的问题，Lin等[18]提出了特征分布对齐技术来增加现有Transformer模型的精度。另外，蒲瞻星等[19]针对小样本下的视频行为识别问题，提出了基于深度特征与流形特征融合的方法；丁静等[20]关注老年人的日常行为识别问题，提出了基于多模态多粒度图卷积网络的方法。

另一方面，对比学习已在计算机视觉领域取得显著进展，并被广泛用于提升图像与视频特征的表征能力。Chen等人提出的SimCLR框架[5]以及He等人提出的MoCo方法[6]是对比学习在图像自监督表征方面的代表性工作，分别通过大规模数据增强和动态字典机制有效提升了特征的一致性和

可分性。这些方法为后续将对比学习扩展到视频任务提供了重要基础。针对视频数据，Zhang等人[7]提出了感知推理对比方法，通过视频间的多模态对比增强感知推理能力；Qian等人[8]构建了时空对比视频表征学习框架，有效捕捉了视频中的局部与全局时序关系；Dave等人[21]提出了TCLR方法，将时间对比学习用于视频特征提取中，显著增强了模型对时序变化的敏感性。此外，Recasens等人[22]和Tschannen等人[23]分别从视角增强与视频中视觉不变性学习的角度，进一步丰富了视频对比学习在自监督场景中的研究。

然而，这些研究多聚焦于无监督或自监督的视频全局特征学习，尚未深入探讨如何在EAR任务中结合未来预测监督与时序区分性进行特征预测。因此，如何将对比学习机制有效嵌入EAR场景下的有监督预测过程中，充分利用预测-观察序列的时间差异信息，仍是值得探索的问题。

2.2　早期行为学习识别方法

完整视频中的行为特征学习研究已取得显著进展，并广泛应用于计算机视觉相关任务。在此基础上，研究者开始将其延伸到EAR这一前沿任务中，即在仅包含整个行为事件早期片段的视频数据中进行特征学习和行为识别。

相较于完整视频行为识别，EAR任务的核心挑战在于如何基于有限的可观察数据预测未来行为演化路径，从而在行为尚未完成时提高预测精度。通常，研究人员更关注于可观察数据少于50%的场景，希望模型能在更早的时刻给出正确的行为识别结果。目前的EAR的主要工作可以划分为三个方向：未来特征学习、时序建模以及模型学习。基于未来特征学习的方法主要是针对EAR任务组中的特征缺少问题，通过基于现有视频片段预测未来行为的特征，可以显著提高EAR的精度。例如，Liu等[24]将输入视频划分为片段集合，并提出动作语义一致性知识网络来挖掘这些片段内的行为信息进而实现特征学习和EAR。Foo等[25]提出了一种新的专家检索和组合网络，其中每个专家模块由不同的卷积网络组成，用于学习不同视频之间的差异，进而形成更加有效的行为表示。Wang等[26]提出了一种新的负元网络利用对比学习策略来缓解由于不可观察片段导致的判别信息不足的问题。

除了未来特征学习，时序建模也是EAR中的一个重要方面。Stergiou等[27]提出了一种基于瓶颈的注意力模型，通过从细到粗多尺度的渐进采样来捕获动作的时序演变表示行为特征。Zheng等[28]提出了一种多模式对抗性知识蒸馏框架，其核心是基于教师-学生网络架构，利用未来未观察到的视频片段来增强可观察视频的表示，并在此基础上预测行为标签。Tai等[29]提出基于时空注意力分解的高阶新型循环网络，以捕获与特定动作相关的时间依赖性。Camporese等人[30]基于动作原型学习，通过在多层次原型空间中捕捉类别先验和时序演化来进行EAR任务的判别优化。与其不同，本文侧重于利用预测未来特征的生成机制，结合正负对比学习策略，显式强化预测特征与未来真实行为特征的一致性，并通过与早期片段的差异性约束进一步增强判别性。

此外，部分研究者从模型学习策略入手，优化EAR任务的训练方法。Xu等[31]提出了一种用于EAR模型简单但有效的训练策略：动态上下文移除。通过动态地安排不同训练阶段上下文的可见性来增加预测难度，直到满足最终的训练目标。Weng等[32]针对现有方法忽略EAR任务中负样本数据存在多样性的问题，在现有的EAR模型中引入了类别排除的策略，以提高模型的准确率。

尽管基于Transformer的行为识别模型在大规模行为识别数据集中展现出卓越性能，但EAR任务中的特征学习仍然面临诸多挑战。例如，时序信息的不完整性和片段间特征的时空不连续性严重影响预测精度。尽管已有研究尝试从时序关系建模与学习策略优化的角度提升EAR模型的表现，当前方法仍主要聚焦于对已观察视频数据的编码，而对于未知片段信息的解码预测仍是该领域的薄弱环节。针对该问题，本文基于交叉注意力机制构建了未来特征生成模型，并基于对比学习对模型进行训练以实现更加准确的未来特征预测与EAR。

3　方　法

本文构建了一种基于对比学习的早期行为识别模型（Contrastive Learning-based Early Action Recognition，CLEAR），其整体模型架构如图3所示。CLEAR模型主要包括视频特征编码模块、未来特征解码网络以及对比学习模块三部分。下文将分别从特征编码、特征解码以及模型的训练与推理过程对CLEAR模型进行详细阐述。

对于给定的训练视频数据X，假设视频共包含t帧，行为标签为Y。首先，以第k帧图像为分界点，

将视频划分为前、后两个片段：即可观察片段$X^{(k)}\in\mathcal{R}^{3\times k\times h\times w}$与不可观察片段$\tilde{X}^{(k)}\in\mathcal{R}^{3\times(t-k)\times h\times w}$其中，$h$与$w$分别表示视频帧图像的高度和宽度。可观察片段$X^{(k)}$在训练和测试阶段均为已知数据，而不可观察片段$X^{(k)}$仅在训练阶段已知，在测试阶段未知。此外，可以定义当前视频观察率$\gamma=\frac{k}{t}$。

基于上述定义，EAR任务的目标可具体描述为：利用给定观察率下的训练数据，训练出一个有效的模型$M(\cdot)$，在输入测试视频$\hat{X}$后，预测其对应的未来行为特征F^*及行为类别概率P^*，具体表达如公式(1)所示。

$$[F^*,P^*]=M\left(\hat{X}\right) \tag{1}$$

3.1 特征编码

在训练过程中，对于给定的可观察视频和不可观察视频片段，首先利用特征骨干网络对其进行特征提取。本文选用Swin Transformer作为骨干网络计算输入视频片段的特征表示，如公式(2)所示。

$$\begin{aligned}F_\gamma&=f_{swin}\left(X^{(k)}\right)\\F_u&=f_{swin}\left(\tilde{X}^{(k)}\right)\end{aligned} \tag{2}$$

其中，$f_{swin}(\cdot)$表示Swin Transformer模型，F_γ和F_u分别表示可观察片段$X^{(k)}$和不可观察片段$\tilde{X}^{(k)}$对应的视频特征。Swin Transformer模型主要由自注意力机制模块堆叠而成，每个模块包含线性映射、多头自注意力、残差连接和前馈神经网络。此外，相邻模块之间通过空间下采样操作实现特征维度的变换。具体而言，在Swin Transformer中，对于大小为$k\times h\times w\times 3$的视频输入，使用尺寸为$2\times4\times4\times3$窗口进行不重叠的滑动采样，得到$\frac{k}{2}\times\frac{h}{4}\times\frac{w}{4}$个视频单元。随后，将每个单元的视频数据展平为一维向量，并通过线性映射层将特征维度转换为c维，从而最终将输入视频转换为尺寸为$\frac{k}{2}\times\frac{h}{4}\times\frac{w}{4}\times c$的特征表示，并送入骨干网络进行深度特征学习。经过特征提取后，输出的视频特征可表示为$F\in\mathcal{R}^{\frac{k}{2}\times d}$，其中$\frac{k}{2}$表示特征序列的长度，$d$表示单个特征的维度，该维度由模型内部的空间下采样和线性映射操作共同决定。

然而，Swin Transformer模型仅对视频片段相邻帧之间的局部空间特征进行编码，并未有效建模视频片段内行为的全局时序依赖关系。因此，本文在Swin Transformer基础上进一步设计了基于注意力机制的时序编码器，用于提取视频帧之间的全局时序行为特征。具体而言，在时序编码器的输入端，本文首先初始化一个可学习的特征单$token\in\mathcal{R}^d$，并将其嵌入到视频特征F中，如公式(3)所示：

$$\tilde{F}_\gamma=Stack(F,token) \tag{3}$$

其中，$token\in\mathcal{R}^d$表示可学习的特征单元；$\tilde{F}_\gamma\in\mathcal{R}^{\left(\frac{k}{2}+1\right)\times d}$表示经过特征单元嵌入后的视频特征。由于时序编码器中缺乏特征序列的位置信息，因此，在执行注意力计算之前，进一步将可学习的位置编码E_{pos}与输入特征$\tilde{F}_\gamma$进行逐元素相加融合，以补充位置信息，如公式(4)所示：

$$X_0=\tilde{F}_\gamma\oplus E_{pos} \tag{4}$$

其中，$E_{pos}\in\mathcal{R}^{\left(\frac{k}{2}+1\right)\times d}$表示可学习的位置嵌入，符号$\oplus$表示逐元素相加操作，$X_0\in\mathcal{R}^{\left(\frac{k}{2}+1\right)\times d}$表示融合位置嵌入后的特征，作为时序编码器的输入。

本文的时序编码器由n个结构相同的自注意力层堆叠而成，每一层均包括多头自注意力模块、残差连接、前馈神经网络以及层归一化操作，如图4所示。具体而言，第i个时序编码层以上一层的输出特征X_{i-1}为输入，其计算过程如公式(5)所示：

$$\begin{aligned}X'_{i-1}&=MSA\left(Norm\left(X_{i-1}\right)\right)+X_{i-1}\\X_i&=FFN\left(Norm\left(X'_{i-1}\right)\right)+X'_{i-1}\end{aligned} \tag{5}$$

其中，$Norm(\cdot)$表示归一化操作，$MSA(\cdot)$表示多头自注意力运算，$FFN(\cdot)$表示前馈网络。在经过n个时序编码层的迭代更新后，输入特征序列最终被更新为$X_n\in\mathcal{R}^{\left(\frac{k}{2}+1\right)\times d}$。最后，将序列中对应初始$token$位置的特征$X_n$分离出来，作为最终的视频行为特征表示，供后续解码网络进一步处理。

3.2 解码网络

在EAR任务中，解码网络的目标是基于已观察到的特征，预测后续尚未观察到的行为特征。为实现这一目标，本文首先定义了l个可学习查询，用以表示待预测的行为特征，记作$Q^0=\left\{Q_1^0,Q_2^0,\cdots,Q_l^0\right\}\in\mathcal{R}^{l\times d}$。随后，将初始查询$Q^0$与编码器输出的特征$token$共同作为解码网络的输入，通过解码网络不断更新可学习查询，最终获得待预测的行为特征。

本文所采用的解码网络架构与原始Transformer的解码器保持一致，由m个结构相同的解码层堆叠构成，如图4所示。每个解码层包含三个核心计算

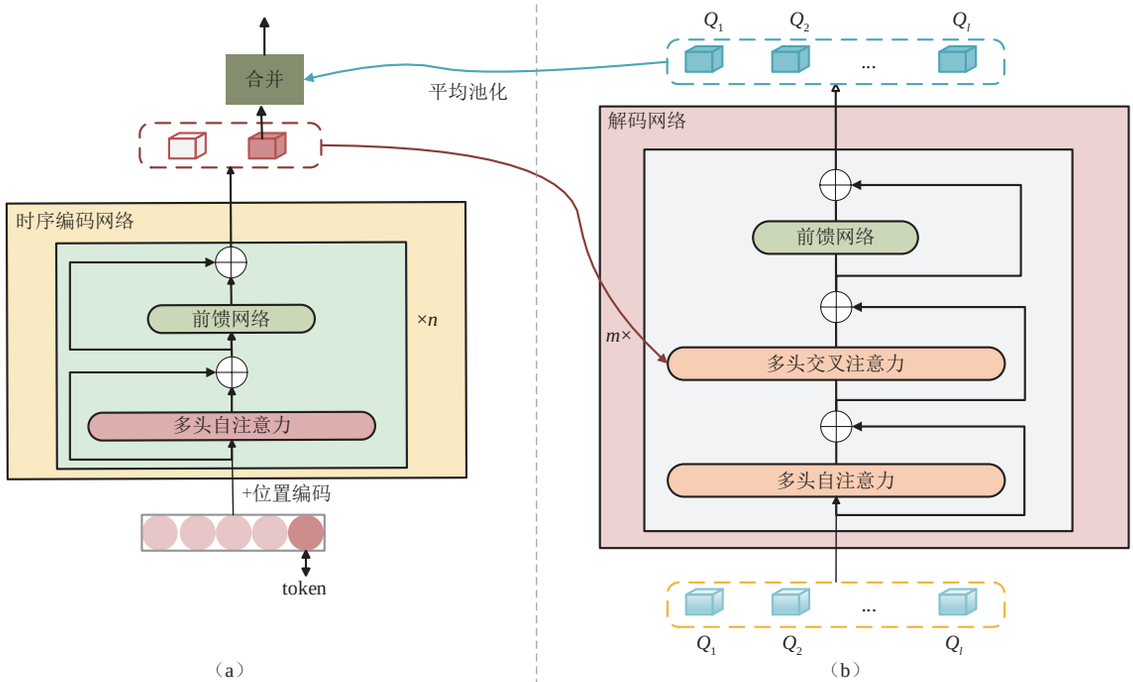

图4　时序编码网络和解码网络示意图

模块：多头自注意力机制、多头交叉注意力机制以及前馈神经网络。同时，各模块均采用残差连接与层归一化进行优化，即执行残差叠加与层归一化处理，从而有效缓解梯度消失问题，提升模型训练的稳定性。每层解码层以上一层更新好后的查询和编码器输出的*token*作为输入，第一层输入的查询为初始查询。首先，通过自注意力操作和残差操作对输入查询进行计算。其次，将得到的查询和编码器的*token*通过交叉注意力进行融合，并融合后的结果与输入查询进行合并，得到更新的查询。最后前馈网络对该查询进行映射，并映射结构和输入进行合并，得到最终输出的查询。解码层的数学表达如公式(6)所示。同时，相较于原始的Transformer解码网络[33]，在本文的方法中采用并行解码的方式对输入查询进行解码。

$$\begin{aligned} H&=MSA\left(Norm\left(Q^{i-1}\right)\right)+Q^{i-1},\\ Q'&=MCA\left(Norm\left(H\right),Norm\left(token\right)\right)+H,\\ Q^{i}&=FFN\left(Norm\left(Q'\right)\right)+Q' \end{aligned} \tag{6}$$

其中，$MCA(\cdot,\cdot)$表示多头交叉注意力机制；Q^{i-1}表示第$i-1$层更新后的查询，也是第i个解码层的输入。首先，可学习查询向量通过自注意力机制进行语义交互与特征增强；更新后的查询与编码网络输出的全局特征进行交叉注意力计算，实现查询与可观察特征的深度融合；最终通过前馈神经网络完成特征空间的非线性映射，输出解码结果。

在本文所提出的方法中，多头交叉注意力机制的实现具体如下：首先，将输入特征分别映射为查询、键和值向量，即将通过自注意力机制增强后的特征H映射为查询向量，而将时序编码网络输出的早期片段特征*token*映射为键和值向量。随后，通过多头交叉注意力机制对上述两类特征进行融合，得到更新后的查询特征H，具体过程如公式(7)所示：

$$\begin{aligned} \boldsymbol{Q}_h&=Norm\left(H\right)\boldsymbol{W}_Q,\\ \boldsymbol{K}_t&=Norm\left(token\right)\boldsymbol{W}_K,\\ V_t&=Norm\left(token\right)\boldsymbol{W}_V,\\ H^j_{mca}&=Softmax\left(\boldsymbol{Q}^j_h\left(\boldsymbol{K}^j_t\right)^{\mathrm{T}}/\sqrt{d_h}\right)V^j_t,\\ \hat{H}_{mca}&=Concat\left(H^1_{mca},H^2_{mca},...,H^J_{mca}\right)W^O_{mca} \end{aligned} \tag{7}$$

其中，$\boldsymbol{Q}_h\in\mathcal{R}^{l\times d}$，$\boldsymbol{K}_t\in\mathcal{R}^{1\times d}$和$V_t\in\mathcal{R}^{1\times d}$分别表示特征映射后的查询、键和值向量；$\boldsymbol{W}_Q\in\mathcal{R}^{d\times d}$，$\boldsymbol{W}_K\in\mathcal{R}^{d\times d}$和$\boldsymbol{W}_V\in\mathcal{R}^{d\times d}$分别是生成对应查询、键和值的线性映射矩阵。公式(7)的前三行计算表达式分别对应对输入的特征进行查询、键和值向量的映射过程；第4行表达式对应查询和键向量点乘计算注意力权重，并与值向量进行映射的过程；最后一个表达式则表示多头注意力机制中，多个注意力头输出的融合过程。在具体实现过程中，本文采用多头注意力机制，即将原始输入特征分解为多个注意力头(假设为J个)，使得模型能够并行关注输入特征的不同维度空间，每个注意力头独立地执行注意

力计算，并从不同的特征子空间中捕获信息。具体而言，输入特征的维度d被均分为J个子空间，每个注意力头的特征维度为$d_h=\frac{d}{J}$。因此，第j个注意力头对应的查询、键和值分别记作$\boldsymbol{Q}_h^j\in\mathcal{R}^{l\times d_h}$、$\boldsymbol{K}_l^j\in\mathcal{R}^{1\times d_h}$、$\boldsymbol{V}_l^j\in\mathcal{R}^{1\times d_h}$；相应的，第$j$个头的输出特征为$H_{mca}^j\in\mathcal{R}^{l\times d_h}$。最后，将各个注意力头的输出通过串联操作进行特征融合，并通过线性映射矩阵$\boldsymbol{W}_{mca}^O\in\mathcal{R}^{d\times d}$映射至原始特征空间，以获得多头交叉注意力机制的最终输出$\hat{H}_{mca}\in\mathcal{R}^{l\times d}$。

经m层的迭代，最后一层的输出$\boldsymbol{Q}^m$通过平均池化操作对每个查询向量进行信息聚合，得到最终预测的不可观察片段行为特征，如公式(8)所示。

$$F^*=avgPooling(\boldsymbol{Q}_1^m,\boldsymbol{Q}_2^m,\cdots,\boldsymbol{Q}_l^m) \tag{8}$$

3.3 模型训练和推理

在对已观察视频片段进行行为特征编码并对不可观察片段进行行为特征预测的基础上，本文设计了三个损失函数用于模型的训练，包括行为分类损失、特征距离损失和对比损失。

首先，行为分类是EAR任务的基础目标。为实现准确的EAR预测，本文将可观察早期片段的行为特征与解码网络预测得到的不可观察片段行为特征进行拼接，构成完整的视频行为特征表示，并基于该融合特征实现行为类别的预测，具体计算如公式(9)所示：

$$\begin{aligned}F_{act}&=Concat(token,F^*)\\P^*&=classifier(F_{act})\end{aligned} \tag{9}$$

其中，$Concat(\cdot,\cdot)$表示特征串联操作，$classifier(\cdot)$表示分类器。本文中所用的分类器为多层感知机(MLP)，并采用softmax函数作为激活函数输出行为类别的概率分布，记作P^*。随后，为了使模型学习到能够区分行为类别的判别性特征，并保证分类分支具备准确的分类能力，本文采用常用的交叉熵损失作为约束，如公式(10)所示：

$$L_{CE}=CE(Y,P^*) \tag{10}$$

其中，Y表示行为标签。

其次，为确保解码网络预测的不可观察片段特征尽可能接近真实的不可观察行为特征，本文设计了特征距离损失，用于约束模型对不可观察片段特征的学习过程。特征距离损失实际上也是对特征相似度的衡量，其计算策略有如欧式距离，余弦相似度等方式。为了和已有的工作[4]进行公平的对比，采用欧式距离作为损失函数。特征距离损失的具体定义如公式(11)所示：

$$L_{fea}=\|F^*-token_u\|_2 \tag{11}$$

其中，$token_u$表示训练阶段利用时序编码器从真实的不可观察视频片段中提取得到的特征，作为不可观察片段特征学习的监督标签。

最后，为进一步提高模型所预测的不可观察片段特征的区分能力和鲁棒性，本文引入对比学习策略来约束特征学习过程。根据EAR任务的设计假设，模型所预测的不可观察片段特征应与真实的不可观察片段特征相近，而与可观察片段的特征存在差异。本文以真实不可观察片段特征$token_u$与预测特征F^*构成正样本对，而以可观察片段的特征序列X_n中各特征与预测特征F^*构成负样本对。在这种设计下，每个训练样本具有1个正样本对和$\left(\frac{k}{2}+1\right)$个负样本对。通过正负样本之间的对比损失对模型进行训练。鉴于本文所设计的正负样本构建策略为多负样本场景，模型训练中采用了信息噪声对比估计(Information Noise Contrastive Estimation，InfoNCE)损失函数进行优化，具体计算过程如公式(12)所示。

$$L_{NCE}=-\log\frac{\exp\left(\left(F^*\cdot token_u\right)/\tau\right)}{\sum_{i=0}^{\left(\frac{k}{2}+1\right)}\exp\left(\left(F^*\cdot X_n^{(i)}\right)/\tau\right)} \tag{12}$$

其中，$\cdot$表示向量的内积运算，τ表示温度参数，用于调节对比损失的敏感度。

综上所述，本文最终所使用的模型训练损失由上述三项损失共同构成，如公式(13)所示：

$$Loss=L_{CE}+L_{fea}+\lambda L_{NCE} \tag{13}$$

其中，λ表示对比损失的权重。本文将通过后续的实验讨论来确定其具体的数值。

与训练阶段早期片段和后期片段均为已知不同，在EAR任务的推理阶段，模型仅能够获得视频的早期片段作为输入。首先，通过骨干网络与时序编码器对早期片段的行为特征进行提取；随后利用解码网络预测与之对应的后期不可观察片段的行为特征；最后再将已观察的早期特征与预测得到的后期特征融合，作为整体行为特征用于最终的行为识别，如图3所示。此外，本文中所引入的特征距离损失和对比学习策略仅用于训练阶段，以增强模型学习的鲁棒性和特征区分能力，而在推理阶段并不会涉及这些额外的损失计算。因此，这些策略的引入并未额外增加模型在推理阶段的计算负担或推理复杂性。

3.4　对比学习在EAR任务中的理论分析

本文引入的对比学习机制可被视为一种基于表示学习的结构优化方法，其目标是在特征空间中拉近预测未来特征与真实后期特征之间的距离，同时扩大其与当前观察段特征之间的差异。这一约束过程与SimCLR[5]中提出的策略具有一致性，本质上提升了特征表示的类间可分性和任务相关性，从而增强了模型对未来行为的判别能力。同时，Oord 等人提出的CPC模型[3]进一步表明，通过构建上下文预测任务并施加对比约束，能够有效提升模型对未来信息的建模能力。本文正负样本对的构建方式在思想上与CPC类似：均通过“当前上下文－未来目标”之间的语义差异学习更加有意义的未来特征。此外，本文的正负样本对不仅体现了语义上的相似与对立关系，还隐含了时间顺序的演化约束，即以“预测特征－真实未来特征”为正样本对、“预测特征-早期片段特征”这个存在时间先后顺序的序列为负样本对，使得模型在学习语义判别特征的同时，也强化了对视频时间顺序的理解与建模。这一顺序感知的对比学习框架，有助于增强模型对行为演化趋势的建模能力，进一步提升早期行为识别任务中的预测准确性。

为进一步解释所提出对比学习机制在EAR任务中提升性能的理论依据，本文从特征表示学习与信息论角度进行分析。首先，根据Oord等人的研究[3]，在对比学习框架下的EAR任务中，InfoNCE损失函数可视为最大化预测特征F^*与真实未来特征$token_u$之间的互信息（Mutual Information，MI）的下界，即

$$I(F^*, token_u) \geqslant \log(N) - L_{NCE} \tag{14}$$

其中，N表示正负样本的数量，在本文的EAR任务中为$\left(\frac{k}{2}+2\right)$。因此，通过最小化InfoNCE损失，模型等价于最大化F^*与真实未来特征$token_u$之间互信息的下界，从信息论角度保证预测的特征包含更多的关于未来行为的信息。这为在EAR框架下对比学习提供了明确的理论依据。

另一方面，在本文的设计中，预测特征还需具备与早期特征的判别性差异，以避免冗余。为了实现这个目标，在本文的设计的InfoNCE函数中，预测特征和早期特征组成了样本对。在InfoNCE的最优化过程中，损失函数倾向于提高正样本对的相似度，并降低各负样本对的相似度；因此，模型可在训练中减少与负样本的混淆，从而在特征空间内形成更清晰的判别边界。综上所述，InfoNCE优化目标本质上在预测特征和未来特征的语义一致性以及预测特征和早期特征差异性之间实现平衡，从理论上解释了本文设计的对比学习机制能有效提升EAR性能的原因。

4　实验结果与分析

4.1　数据集和实验设置

数据集：本文采用Something-Something V2（SSv2）和UCF101数据集来验证所提出EAR方法的有效性。SSv2数据集则包含了约22万个视频样本，涵盖174个细粒度的人与物体交互动作类别。由于其视频数量庞大且类别丰富，SSv2具有较大挑战，被广泛应用于更大更复杂网络模型在视频动作识别任务的性能评估。UCF101数据集由13 320个视频组成，涵盖101个动作类别，全部视频均采集自YouTube。每个动作类别进一步细分为25个组别，每组包含4至7个动作视频。根据动作特征，该数据集可以归为五种类型：人与物体交互、纯身体运动、人与人交互、演奏乐器以及体育运动。UCF101也是人体行为理解领域研究中常用的标准数据集之一，官方提供了三种不同的训练与测试划分方式分别命名为Split01、Split02和Split03，并采用平均准确率作为评测指标。

评价指标：与现有EAR相关研究一致，本文选取在不同观察率条件下的动作分类准确率作为性能评价指标。由于EAR任务更加关注低观测数据场景，因此大部分现有方法通常在观察率为10%至50%的范围内进行性能评估。本文亦遵循这一常规设置，对所提出方法的有效性进行评测与分析。

实验设置：本文所提出的方法在搭载Intel i7-9700 CPU、64 GB内存及NVIDIA RTX 3090 GPU的计算平台进行实验。在视频序列处理方面，本研究采用和现有工作同样的连续采样策略[27,34-35]：即随机确定每个视频序列的起始帧索引，并从该起始帧开始连续采样16帧。当观测序列长度不足16帧时（例如观察率为10%的情况），则通过从视频序列开头重复采样帧的方式，将序列长度补齐至16帧。特征提取骨干网络采用在Kinetics-400数据集上预训练的Swin-S模型权重。训练过程中，网络参数的优化采用AdamW优化器，其权重衰减系数设为1e－3。在UCF101数据集上，训练批次大小设置为16；

在Something-Something V2数据集上，考虑到更高的计算开销，批次大小调整为8。两组实验的训练轮次均设置为60。在模型参数设置方面，编码器与解码器的注意力头数分别设置为8和4。此外，编码网络中的自注意力模块层数n设置为5层，解码器中的交叉注意力模块m层数则设置为6层。在接下来的实验中，本文也将对这些关键参数的设置及其对模型的影响进行讨论。

4.2 与现有方法的对比分析

为了评估所提出方法的有效性，本文将其与现有主要的EAR方法在SSv2和UCF101数据集上进行了实验对比分析。首先，本文在更大规模且更具挑战的数据集SSv2上验证所提出方法的有效性，实验结果如表1所示。与Liu等人[24]的方法相比，本文方法在观察率ρ=20%和ρ=40%条件下的识别准确率分别提高了6.12%和7.76%，平均提高了6.94%。与AAttNet[4]方法相比，在观察率为10%、20%、30%、40%和50%的情形下，本文方法的识别精度分别提高了3.81%、3.63%、4.44%、5.27%和5.69%，平均提高了4.96%。此外，从模型参数量来看，本文提出的CLEAR方法仅为136.1 M，远低于AAttNet的364.3 M，同时显著优于Swin-B(88 M)在多项观察率下的识别性能。这表明本文方法在有效提升EAR任务识别准确率的同时，还具有较优地控制了模型的复杂度，充分体现了良好的效率－性能平衡，进一步验证了其在实际部署场景下的应用潜力。

其次，表2中展示了本文的方法与其他方法在UCF101数据集上的对比结果。实验结果表明，与现有的经典EAR模型(如MSRNN[35]、Transfer[36]、T-S Model[37]、DBDNet[38]、S-T Relation[39]、DRL[34]和AmwmNet[40])相比，本文提出的方法在所有观察率下均表现出一致的准确率优势，其平均识别准确率分别提升了5.72%、4.05%、2.02%、2.43%、6.25%、3.08%和2.14%。此外，近期所发表的工作，如与Liu等人[24]的方法相比，在观察率ρ分别为10%、20%和30%时，本文方法的准确率分别提高了0.5%、1.79%和0.97%，平均提高了约1.09%。与AAttNet[4]方法相比，在观察率为10%、20%、30%、40%和50%时，本文方法的准确率分别提升了0.8%、0.59%、0.8%、0.35%和0.41%，平均准确率则由89.44提升至90.03，提高了0.59%。这些实验结果进一步验证了本文方法在EAR任务中的有效性。

表1 所提出方法在SSv2数据集上的对比结果

方法	参数量	数据	观察率					
			10%	20%	30%	40%	50%	平均
Swin-B[27]	88 M	RGB视频	14.4	-	23.5	-	31.5	-
Liu[24]	-		-	16.21	-	23.67	-	-
AAttNet[4]	364.3 M		14.98	18.70	22.07	26.16	31.42	22.27
CLEAR	136.1 M		18.79	22.33	26.51	31.43	37.11	27.23

然而，相较于当前利用人体骨架数据的EAR模型，本文方法的识别精度仍有一定的差距。其主要原因在于人体骨架数据能够通过人体关节点的时序变化更加准确且高效地对人体动作进行建模，从而减少视频背景的干扰。在未来的研究工作中，本文将进一步探索融合多模态数据的EAR模型，以提升模型的性能表现。

此外，本文在UCF101数据集上的实验结果显示，当观察率从10%提升至50%时，模型识别准确率的增幅相对有限，这一现象与UCF101数据集本身的特性密切相关。UCF101包含大量具有显著场景先验和明显初始动作线索的行为类别，例如跳水、击剑等，使模型能够在极早期就提取到足够的判别信息，从而导致随着观察率增加所带来的信息增量对最终分类结果的提升相对较小。此外，UCF101上的整体分类准确率已超过88%，也反映出该数据集在部分行为类别上的区分难度相对较低。相比之下，本文在更具挑战性的SSv2数据集上同样进行了验证。SSv2作为细粒度动作数据集，包含大量需要依赖动作演化过程才能进行准确判别的交互行为，显著增加了任务的复杂性。实验结果表明，本文方法在该数据集上相比现有方法平均识别准确率提升了4.96%，进一步验证了所提出方法在处理细粒度、依赖时序信息的复杂行为场景中的有效性。

4.3 消融实验

在本节中将通过消融实验验证所提出方法中关键模块的有效性。鉴于本文方法在UCF101数据集上的性能提升相较于在SSv2数据集上不够显著，本

表2　UCF101数据集上的对比结果

方法	数据	观察率					
		10%	20%	30%	40%	50%	平均
MSRNN[35]	RGB视频	68.00	87.39	88.16	88.79	89.24	84.31
Transfer[36]		80.00	84.70	86.90	88.60	89.70	85.98
T-S Model[37]		83.32	87.13	88.92	89.92	90.85	88.01
DBDNet[38]		82.70	86.60	88.30	89.70	90.60	87.60
S-T Relation[39]		80.24	-	84.55	-	86.28	83.69*
DRL[34]		83.90	85.10	86.90	88.37	90.50	86.95
AmwmNet-RGB[40]		85.95	87.47	88.21	88.57	89.26	87.89
TemPr[27]		85.70	91.40	92.10	92.70	93.50	91.08
Liu[24]		87.56	87.59	89.38	-	-	-
AAttNet[4]		87.26	88.79	89.55	90.58	91.00	89.27*/89.44
HARDer-Net[41]	人体骨架序列	87.26	-	92.65	-	94.32	91.63
ERA-Net[25]	RGB视频 + 人体骨架序列	89.14	-	92.39	-	94.29	91.94
CLEAR(Split01)	RGB视频	87.42	88.66	89.61	90.40	90.75	89.26*/89.37
CLEAR(Split02)		88.40	90.01	91.14	91.94	92.11	90.55*/90.72
CLEAR(Split03)		88.37	89.48	90.29	90.45	91.37	90.01*/89.99
CLEAR(AVG)		88.06	89.38	90.35	90.93	91.41	89.94*/90.03

注："*"表示0.1、0.3和0.5观察率下的平均准确率；"-"表示原始论文中没有提供对应的结果。

文选择在UCF101数据集上开展更为细致的消融实验。通过对关键模块与超参数的系统性分析，旨在进一步探讨所提出方法在UCF101数据集上的表现特性，从而进一步验证其有效性。本文后续各节中的消融实验均基于UCF101数据集的第一种官方训练/测试划分(Split01)进行。

为了验证所提出对比学习方法的有效性，本文设计了一组消融实验，将训练过程中采用对比损失的模型与未引入对比损失的模型进行对比分析。实验结果如表3所示。从结果可见，在不同观察率条件下，采用对比学习方法的模型均实现了预测精度的提升。具体而言，采用对比学习的模型在平均预测准确率上相比未使用对比学习的模型平均准确率从88.64%提升到了89.37%，提高了0.73%。上述实验结果表明，本文提出的对比学习方法能够有效增强模型的特征判别能力，通过提高特征空间的类间区分度，使模型能够更早且更准确地实现动作类别识别。具体地，通过对比损失构造负样本对将预测的未来特征与早期特征显式拉开，可迫使模型学习到更多与未来行为相关的判别特征，避免生成的预测特征过度接近已观察特征而失去额外信息增益。

另外，考虑到特征距离损失与对比学习损失在计算预测特征与正样本之间相似度时存在一定的重叠，本文进一步设计了特征距离损失的消融实验，以验证其在整体损失框架中的必要性。实验结果表明，当在总损失中移除特征距离损失后，模型的平均识别精度由89.37%下降至87.81%，降低了1.56%。这一显著下降说明，仅依赖对比学习所形成的正负样本分布拉开还不足以充分捕捉预测特征与真实未来特征之间的精确对应关系。对比学习更偏向于优化全局特征空间中类别间的判别性，而特征距离损失则直接最小化预测特征与真实未来特征的欧氏距离，从细粒度层面增强了两者的一致性，有助于提高未来特征预测的准确度。因此，利用特征距离损失对模型进行额外约束，能够有效弥补对比学习在局部特征对齐上的不足，是提升模型性能的重要因素。

表3　关于损失函数的消融实验

损失函数	观察率					
	10%	20%	30%	40%	50%	平均
总损失	87.42	88.66	89.61	90.40	90.75	89.37
无对比学习损失	87.34	88.06	88.53	89.37	89.90	88.64
无特征距离损失	84.85	87.47	88.55	88.55	89.61	87.81

表4进一步分析了对比损失中超参数τ对模型然而，实验也发现进一步减小τ反而会导致模型精度下降。这是由于过小的τ会导致对比学习中的相似性分布过于集中，使模型过度关注正样本而忽略负样本之间的有效区分，从而难以充分学习数据中蕴含的复杂关系，最终导致性能下降。此外，本文的训练过程中共采用三个损失函数对模型进行联合优化。其中，分类损失与距离损失为EAR任务中的常

规损失函数，而对比损失则由本文提出并引入。为了更好地平衡不同损失函数的贡献，本研究进一步探讨了对比损失的权重系数λ对模型性能的影响，实验结果如表5所示。当$\lambda=0.5$时，模型达到最优性能，平均准确率为89.37%。与λ分别取值为1、0.7、0.3和0.1的情形相比，其平均准确率分别提升了0.28%、0.34%、0.11%和0.16%。

表4 对比损失中超参数的对比实验

超参数τ	观察率					
	10%	20%	30%	40%	50%	平均
0.08	87.23	88.61	89.16	89.93	90.56	89.10
0.09	87.42	88.66	89.61	90.40	90.75	89.37
0.1	87.39	88.55	89.11	89.90	90.62	89.11

表5 对比损失的权重系数对比实验

λ	观察率					
	10%	20%	30%	40%	50%	平均
1	87.15	88.61	89.27	89.96	90.48	89.09
0.7	87.10	88.37	89.32	89.93	90.43	89.03
0.5	87.42	88.66	89.61	90.40	90.75	89.37
0.3	87.66	88.66	89.21	89.96	90.80	89.26
0.1	87.81	88.53	88.90	90.25	90.59	89.21

为了进一步分析编码器层数对模型动作识别性能的影响，本文设计并实施了针对不同编码器层数的对比实验，相关实验结果如表6所示。从表中可以观察到，5层编码器结构在识别性能上表现最优。具体而言，相较于采用4层与6层编码器结构的模型，5层编码器结构的平均识别准确率分别提高了0.44%和0.27%。值得注意的是，这一性能优势在所有观察率条件下均表现稳定，因此使用5层编码器结构对提升模型整体性能有显著的有效性。

表6 编码器层数对模型识别性能影响的对比实验结果

编码器层数	观察率					
	10%	20%	30%	40%	50%	平均
4	86.86	88.08	89.56	89.90	90.27	88.93
5	87.42	88.66	89.61	90.40	90.75	89.37
6	87.34	88.47	89.58	90.01	90.11	89.10

此外，本文进一步评估了解码器层数对模型性能的影响，并开展了相应的实验分析，实验结果如表7所示。从表中可以看出，在不同观察率条件下，当解码器层数为6时，模型的性能表现优于其他层数设置。特别是在观察率为10%时，6层解码器结构的识别准确率相较于5层与7层的模型分别提升了0.56%和0.43%。在整体平均准确率方面，6层解码器结构的表现也分别比5层与7层模型提高了0.37%与0.42%。基于以上实验结果，本研究提出的方法最终选择采用6层的解码器结构。

表7 解码器层数对模型识别性能影响的对比实验结果

解码器层数	观察率					
	10%	20%	30%	40%	50%	平均
5	86.86	88.32	89.11	90.22	90.48	89.00
6	87.42	88.66	89.61	90.40	90.75	89.37
7	86.99	88.18	89.51	89.88	90.19	88.95

基于交叉注意力的未来特征生成模块是本文所提出方法的关键组成部分之一。为进一步探讨该模块对模型识别性能的影响，本文设计并实施了针对解码器中不同注意力头数量的对比实验，具体结果如表8所示。从实验结果可知，当注意力头的数量设置为4时，模型取得最优性能表现，其平均准确率相较于注意力头数量设置为2和6的情形分别提高了0.28%和0.29%。首先，注意力头数量增加会导致模型复杂性上升，可能导致模型在训练数据上的过拟合，从而降低泛化性能。此外，过多的注意力头会显著增加模型参数数量，引入额外噪声降低训练的稳定性。因此，基于实验验证结果，本文最终选择在解码器中采用4个交叉注意力头，以达到最佳的性能效果。

表8 交叉注意力头数量对模型识别性能影响的对比实验

交叉注意力头数	观察率					
	10%	20%	30%	40%	50%	平均
2	87.34	88.34	89.43	89.93	90.40	89.09
4	87.42	88.66	89.61	90.40	90.75	89.37
6	87.18	88.45	89.27	90.11	90.40	89.08

最后，为了探究骨干特征网络对模型性能的影响，本文在UCF101和SSv2两个数据集上分别采用Swin-T和Swin-S作为特征提取骨干，进行了对比实验。实验结果如表9所示。可以看出，使用Swin-S作为骨干网络时，模型在UCF101和SSv2数据集上的平均识别精度分别比采用Swin-T提高了0.12%和0.47%。这一结果表明，选用更强的特征骨干能够有效提升模型的识别性能。此外，本文所提出的网络框架在设计上具备良好的可扩展性，骨干特征提取模块可以灵活替换为更高效的网络结构，以适应后续研究中的进一步优化需求。

4.4 定性分析

为了进一步定性地评估本文提出的CLEAR方

表9　不同特征骨干网络对模型识别性能影响的对比实验

数据集	特征	观察率					
		10%	20%	30%	40%	50%	平均
UCF101	Swin-T	87.32	88.67	89.53	90.15	90.58	89.25
	Swin-S	87.42	88.66	89.61	90.40	90.75	89.37
SSv2	Swin-T	18.37	22.17	26.18	30.54	36.45	26.76
	Swin-S	18.79	22.33	26.51	31.43	37.11	27.23

法的有效性，本节对部分样本的识别结果进行了可视化分析。图5展示了CLEAR方法与AAttNet[4]方法在观察率$\gamma=20\%$条件下的识别性能对比。图中划线标签表示视频序列的真实类别，绿色字体表示模型Top-1预测正确，红色字体则表示模型预测错误。此外，每组结果列出了两种模型预测概率排名前3的类别。

从图中第一组视频片段可以观察到，CLEAR方法准确识别出真实类别“Twisting（wringing）something wet until water comes out”，且所预测的其他候选类别（如“Pretending or trying and failing to twist something”与“Twisting something”）均与动作的语义高度相关，这体现出CLEAR方法对动作语义的准确理解。相较之下，AAttNet法虽然也给出了正确的Top-1预测类别，但其预测的其他类别（如“Squeezing something”）与真实动作存在一定语义差距，体现出AAttNet方法在细粒度动作区分上的不足。第二组视频片段进一步体现了CLEAR方法对精细动作特征的敏锐捕捉能力。CLEAR方法不仅正确预测了真实类别“Something falling like a feather or paper”，而且其后续预测的类别（如tin the air and letting it fall”）也均与视频中的真实动作密切相关，反映出CLEAR方法对动作细节和动态信息的有效建模。相比之下，AAttNet方法虽然也将正确类别排在首位，但后续预测的其他候选类别（例如

图5　CLEAR方法与AAttNet[4]方法在SSv2数据集上的结果可视化

“Moving something towards the camera”)则与真实动作存在明显差异,表现出其对动作的动态特征捕捉不足,且对动作场景理解较为模糊。

在第三组视频序列中,AAttNet方法未能准确识别真实动作类别,而CLEAR方法则准确预测了真实动作类别“Putting something that cannot actually stand upright upright on the table, so it falls on its side”。这表明CLEAR方法能够更加有效地捕获视频中的细微时序特征,从而在存在细节差异的动作识别任务中取得优势。这些分析结果进一步说明,CLEAR方法相比于AAttNet方法,在面对类别间差异较小、动作细节更复杂的SSv2数据集时,具备更高的准确性与更强的鲁棒性。这也验证了CLEAR方法在早期行为识别任务中具备显著的性能提升,能够有效处理真实场景下的复杂视频数据。以上分析结果表明,相比于AAttNet方法,CLEAR方法在细粒度动作识别任务中能够更准确地捕获视频中的动态信息与细微特征,表现出更佳的识别效果与更强的鲁棒性。这一优势进一步验证了CLEAR方法在实际复杂视频场景下的有效性。

为了进一步验证所提出对比学习策略在不同行为类别上的适用性与有效性,本文在观察率为0.1这一极低信息条件下,对所有类别分别统计了使用与未使用对比学习的识别准确率,并筛选出两种方法识别精度差异超过0.04的类别共计35类进行可视化分析,如图6所示。从图中可以观察到,对于如“GolfSwing”、“CleanAndJerk”、“Skiing”、“ParallelBars”、“Taichi”等具有显著动作阶段性或转换特征的行为类别,采用对比学习后其识别准确率相较于未使用对比学习显著提高。这表明在此类行为中,预测特征与早期观察特征之间存在较大的差异,所提出的正负样本构建策略能够充分发挥其区分优势。相反,对于“ApplyEyeMakeup”、“WalkingWithDog”、“HandstandWalking”、“MoppingFloor”等连续性较强、特征变化平滑的行为类别,使用对比学习后准确率出现小幅下降。这与预期一致,即对于这类行为,早期片段特征与未来特征的差异不明显,对比损失的约束作用不如在阶段性显著行为中明显。该实验进一步细化了对比学习在EAR任务中适用的行为场景,同时也揭示了其在处理特征平滑、时序变化缓慢行为时的局限性。未来的研究将针对此类连续性较强的行为,探索更加灵活的正负样本构造策略及自适应机制,以进一步提高模型在多样化场景下的泛化性能。

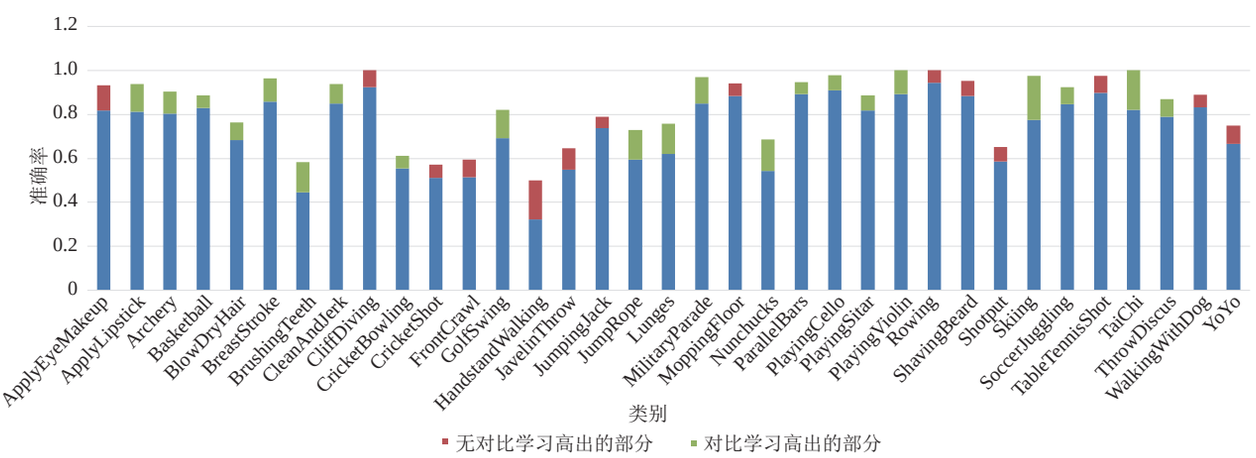

图6 对比学习和无对比学习的方法在具体类别上的识别精度对比

为进一步探究所提出对比学习策略对未来特征空间分布的影响,本文在UCF101数据集上,分别在使用与未使用对比学习的条件下,对预测的未来特征进行了t-SNE可视化,并计算轮廓系数与Davies-Bouldin(DB)指数以量化评估特征空间的聚类情况,结果图如7所示。实验结果表明,未使用对比学习时轮廓系数为0.8500,DB指数为0.2434;引入对比学习后分别为轮廓系数为0.8503,DB指数为0.2411。两组指标整体接近,均显示预测的未来特征已表现出良好的类内紧密性及类间分离性。尽管t-SNE的聚类指标未显示出显著差异,但在UCF101上的分类实验结果显示,引入对比学习后,平均分类准确率从88.64%进一步提升至89.37%。这说明所提出的对比学习策略主要通过优化特征分布的判别边界来增强未来特征的可分性,从而有效提高了EAR任务的预测性能。

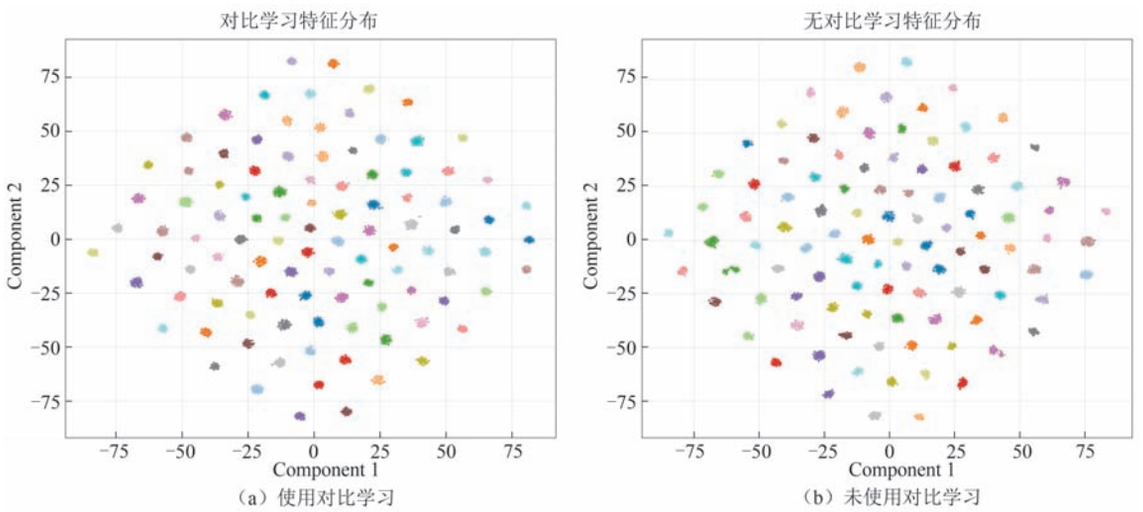

图7　预测未来特征的t-SNE图

为更直观地展现所提出方法在不同观察率下对视频关键区域的关注情况，本文对预测结果进行了类激活图可视化，如图8所示。图中选取了“ApplyEyeMakeup”、“ApplyLipstick”、“CricketShot”以及“Basketball”等行为类别，在观察率从0.1至0.5逐步增加时，观察模型所聚焦的特征区域的变化。从可视化结果可见，在较低观察率(0.1、0.2、0.3)时，模型关注的区域较为分散，热力图覆盖运动的无关区域甚至于背景区域。随着可观察片段逐渐增多，注意力区域逐步收缩至与行为强相关的局部区域，如“ApplyEyeMakeup”逐步聚焦至眼部、“ApplyLipstick”聚焦至嘴唇区域，“CricketShot”集中在挥棒动作，而“Basketball”聚焦至投篮动作和篮球区域。这表明所提出的未来特征预测机制能够在信息逐步补充的过程中有效提升对行为判别性区域的关注能力。

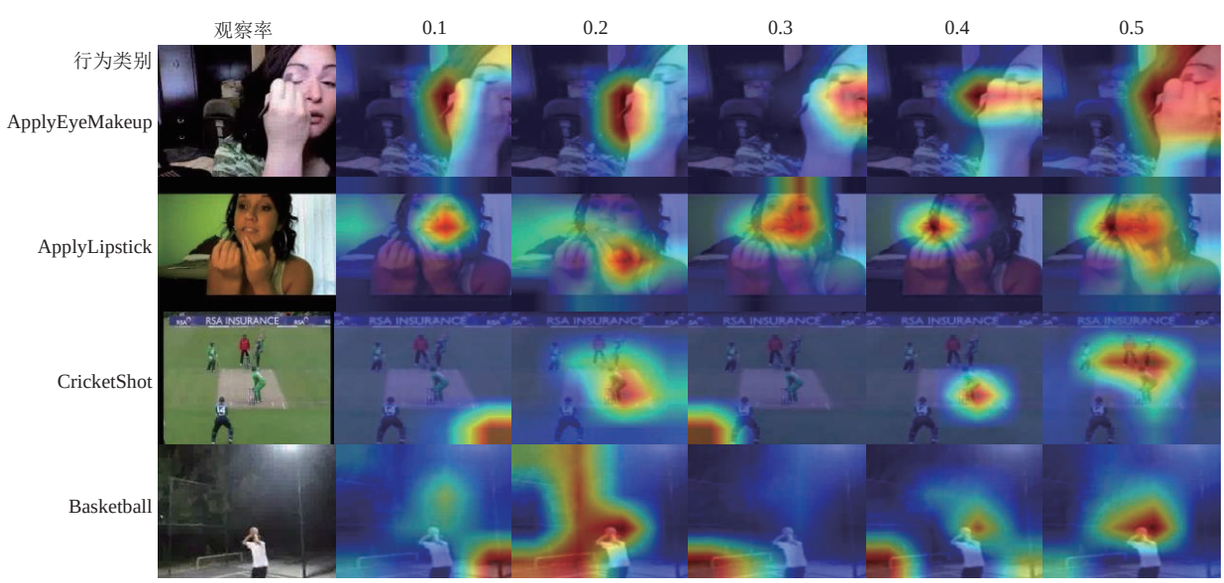

图8　不同观察率的模型类激活图可视化

4.5　实际应用场景的适用性和部署可行性分析

面向实际应用场景与工程部署，基于前文实验与本文方法的特性，本文对所提出方法在实际场景中的适用性与部署可行性进行分析。与常规的行为理解模型在公共安全预警、体育训练分析等系统中的部署一致，本文模型在实际部署时亦采用滑动窗口/重叠采样的方式进行流式推理。系统的端到端时延主要由“采样累计时间＋前向推理时间＋轻量后处理”构成。

首先，尽管本文在训练阶段引入了对比学习与特征距离损失以优化EAR任务，但正如第3.3节

“模型训练和推理”所述，上述损失仅用于训练阶段，推理阶段不包含相应损失项与计算。因此，在实际应用场景中，本文模型的整体架构与主流视觉系统的通用架构一致，即由特征提取—编码网络—解码网络—分类器构成。在此基础上，本文模型在推理部署与系统集成层面具有明确的可实施性与可优化性：一是便于在ONNX/TensorRT等通用引擎中进行算子融合、常量折叠与精度模式切换；二是易于与既有视觉识别系统实现模块化对接与替换；三是便于在实际场景数据上开展迁移学习与小样本微调（如冻结骨干、微调解码器与分类器等），在保持推理计算不变的前提下快速完成场景适配，从而保障部署的可行性与稳定性。

其次，本文方法在参数规模与识别性能之间取得了较优平衡：CLEAR参数量为136.1 M，显著低于AAttNet的364.3 M，并在多种观察率下获得更优准确率，为资源受限场景提供了部署空间。进一步地，特征提取骨干可在Swin-S（约50 M参数）与更轻量的Swin-T（约28 M参数）之间灵活切换。表9显示，将骨干替换为Swin-T后，UCF101与SSv2的平均精度仅出现轻微变化，而整体参数量与算力需求进一步下降；这表明在保持总体性能的同时，可依据算力与延迟约束选择更轻量的骨干，实现低资源占用。综合来看，模型具有推理阶段计算简洁、对骨干选择低敏感、支持实际场景微调等特点，可在不同应用场景上实现快速适配与可持续优化。

5 总 结

本文基于Transformer架构提出了一种基于对比学习的早期行为识别方法。在本文所提出的方法中，使用编码网络对输入的视频进行特征表示，利用解码网络实现未来不可观察片段的特征预测。为了对模型进行更加有效的训练，在本文的方法中引入了对比学习方法。具体来说，将早期视频片段和预测的未来特征组成负样本，而预测的未来特征与真实的特征标签组成正样本。通过对比损失引导模型学习更具差异化的未来特征，从而有效补充早期特征缺失的行为信息，丰富整体特征表达。在消融实验中，本文对比学习策略的有效性，模型参数的影响进行了详细的讨论和验证。同时，与现有方法在UCF101和SSv2数据集上进行了对比实验进一步验证了所提出的方法的有效性。

尽管本文所提出的方法在识别准确率方面取得了一定优势，但仍存在一些不足。从表2的实验结果可以看出，与采用骨架数据的EAR方法相比，本文方法在识别精度上仍有差距，这主要由于基于骨架数据的方法能够更准确、高效地捕捉人体动作的时序变化特征。这一观察为后续研究提供了重要启发，即在未来工作中将进一步探索融合多模态数据的EAR方法，特别是借鉴Alayrac等人[42]提出的多模态对比学习框架，结合RGB视频与骨架信息，以期显著提升EAR任务的识别精度与鲁棒性。

此外，本文方法目前仅在公开数据集上进行了验证，对于满足实际应用中对实时性的需求仍需进一步研究。未来将重点围绕模型推理效率的优化及高效在线推理展开，并结合实际场景开展有针对性的模型优化。同时，考虑到真实环境中可能出现的噪声干扰及数据差异，后续工作还将引入对抗训练和域适应等技术，并且进一步深入探索对比学习策略在EAR任务的变体，以提升模型的识别精度、泛化能力和稳健性。

最后，在文本的实验中亦观察到，所构建的模型在处理连续性较强的行为时效果相对有限。此类行为通常具有阶段过渡不明显、帧间差异微弱、背景先验稳定等特征；在低观察率条件下，早期片段与后续片段在特征空间中的分布往往趋近，从而减弱了本文基于“预测未来特征并区分已观测特征”的学习目标所能利用的判别信息。这个问题影响了本文方法的普适性。针对这个局限，后续工作拟从两方面推进：(1)引入多尺度与可变步长的时序建模，扩展有效时间感受野以捕获缓慢动态与节奏变化；(2)构建阶段感知的局部对齐与片段级对比目标，将全局对比细化为短片段与其未来邻域的对应关系，以提升弱过渡场景下的可分性与稳定性。上述方案均在不改变主体框架的前提下实施，旨在进一步提升对连续性行为的早期判别能力与稳健性，并作为本研究的重点后续方向。

参 考 文 献

[1] Zhang Tian-Yu, Min Wei-Qing, Han Xin-Yang, et al. A survey on future action anticipation in videos. Chinese Journal of Computers, 2023, 46(6): 1315-1338 (in Chinese)
(张天予，闵巍庆，韩鑫阳，等. 视频中的未来动作预测研究综述. 计算机学报, 2023, 46(6): 1315-1338)

[2] Li Zhe, Zhang Hong-Bo, Zhang Miao-Hui, et al. Late feature supplement network for early action prediction. Image and Vision Computing, 2022, 125:104519

[3] Oord Aaron van den, Li Yazhe, OriolVinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv: 1807.03748, 2018

[4] Zhang Hong-Bo, Pan Wei-Xiang, Du Ji-Xiang, et al. Adversarial attention networks for early action recognition. IEEE Transactions on Emerging Topics in Computational Intelligence, 2025, 9(2): 1581-1594

[5] Ting Chen, KornblithSimon, Norouzi Mohammad, Hinton Geoffrey. A simple framework for contrastive learning of visual representations//Proceedings of the International Conference on Machine Learning. Vienna, Austria, 2020: 1597-1607

[6] He Kaiming, Fan Haoqi, Wu Yuxin, et al. Momentum contrast for unsupervised visual representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 9726-9735

[7] Zhang Chi, Jia Baoxiong, Gao Feng, et al. Learning perceptual inference by contrasting//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada, 2019: 1-13

[8] Qian Rui, Meng Tianjian, Gong Boqing, et al. Spatiotemporal contrastive video representation learning//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA, 2021: 6960-6970

[9] Khurram Soomro, Zamir Amir Roshan, MubarakShah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012

[10] Raghav Goyal, Ebrahimi Kahou Samira, VincentMichalski, et al. The "something something" video database for learning and evaluating visual common sense//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5842-5850

[11] Joao Carreira, Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 6299-6308

[12] Cob-Parro Antonio Carlos, Cristina Losada Gutiérrez, Marta Marrón Romera, et al. A new framework for deep learning video based human action recognition on the edge. Expert Systems with Applications, 2024, 238: 122220

[13] Anurag Arnab, Mostafa Dehghani, Georg Heigold, et al. Vivit: A video vision transformer//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 6836-6846

[14] Qing Zhiwu, Zhang Shiwei, Huang Ziyuan, et al. Mar: Masked autoencoders for efficient action recognition. IEEE Transactions on Multimedia, 2023, 26: 218-233

[15] Mou Yuting, Jiang Xinghao, Xu Ke, et al. Compressed video action recognition with dual-stream and dual-modal transformer. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 34(5): 3299-3312

[16] Yang Taojiannan, Zhu Yi, Xie Yusheng, et al. Aim: Adapting image models for efficient video action recognition// Proceedings of the International Conference on Learning Representations. Kigali, Rwanda, 2023: 1-18

[17] Wu Pei-Chen, Li Wen-Bin, Guo Fang, et al. Video anomaly detection based on feature enhancement and modal interaction. Journal of Computer-Aided Design & Computer Graphics, 2024, 37(3): 407-413 (in Chinese)
(吴沛宸，李文斌，郭放，等. 基于特征增强和模态交互的视频异常行为检测. 计算机辅助设计与图形学学报，2024, 37(3): 407-413)

[18] Lin Wei, Mirza Muhammad Jehanzeb, MateuszKozinski, et al. Video test-time adaptation for action recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 22952-22961

[19] Pu Zhan-Xing, Ge Yong-Xin. Few-shot action recognition in video based on multi-feature fusion. Chinese Journal of Computers, 2023, 46(3): 594-608 (in Chinese)
(蒲瞻星，葛永新. 基于多特征融合的小样本视频行为识别算法. 计算机学报，2023, 46(3):594-608)

[20] Ding Jing, Shu Xiang-Bo, Huang Peng, et al. Multimodal and multi-granularity graph convolutional networks for elderly daily activity recognition. Journal of Software, 2023, 34(5): 2350-2364 (in Chinese)
(丁静，舒祥波，黄捧，等. 基于多模态多粒度图卷积网络的老年人日常行为识别. 软件学报，2023, 34(5): 2350-2364)

[21] Ishan Dave, Rohit Gupta, Rizve Mamshad Nayeem, et al. Tclr: Temporal contrastive learning for video representation. Computer Vision and Image Understanding, 2022, 219: 103406

[22] Recasens A, Luc P, Alayrac J B, et al. Broaden your views for self-supervised video learning//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada, 2021: 1255-1265

[23] Michael Tschannen, Josip Djolonga, Marvin Ritter, et al. Self-supervised learning of video-induced visual invariances// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 13806-13815

[24] Liu Xiaoli, Yin Jianqin, Guo Di, et al. Rich action-semantic consistent knowledge for early action prediction. IEEE Transactions on Image Processing, 2023, 33: 479-492

[25] Lin GengFoo, Li Tianjiao, Hossein Rahmani, et al. Era: Expert retrieval and assembly for early action prediction// Proceedings of the European Conference on Computer Vision. Tel Aviv, Israel, 2022: 670-688

[26] Wang Wenqian, Chang Faliang, Zhang Junhao, et al. Magi-Net: Meta negative network for early activity prediction. IEEE Transactions on Image Processing, 2023, 32: 3254-3265

[27] Alexandros Stergiou, Dima Damen. The wisdom of crowds: Temporal progressive attention for early action prediction// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada, 2023: 14709-14719

[28] Zheng Na, Song Xuemeng, Su Tianyu, et al. Egocentric early action prediction via adversarial knowledge distillation. ACM Transactions on Multimedia Computing, Communications and Applications, 2023, 19(2): 59

[29] Tai Tsung Ming, GiuseppeFiameni, Lee Cheng Kuang, et al. Higher-order recurrent network with space-time attention for video early action recognition//Proceedings of the IEEE

International Conference on Image Processing. Bordeaux, France, 2022: 1631-1635

[30] Guglielmo Camporese, Alessandro Bergamo, Lin Xunyu, et al. Early Action Recognition with Action Prototypes. arXiv preprint arXiv:2312.06598, 2023

[31] Xu Xinyu, Li Yong-Lu, Lu Cewu. Dynamic context removal: A general training strategy for robust models on video action predictive tasks. International Journal of Computer Vision, 2023, 131(12): 3272-3288

[32] Weng Junwu, Jiang Xudong, Zheng Wei-Long, et al. Early action recognition with category exclusion using policy-based reinforcement learning. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(12): 4626-4638

[33] Ashish Vaswani, Noam Shazeer, Parmar Niki, et al. Attention is all you need//Proceedings of the Advances in Neural Information Processing Systems. Long Beach, USA, 2017, 30.

[34] HareeshDevarakonda, SnehasisMukherjee. Early prediction of human action by deep reinforcement learning//Proceedings of the 2021 National Conference on Communications. Kanpur, India, 2021: 1-6

[35] Hu Jian-Fang, Zheng Wei-Shi, Ma Lianyang, et al. Early action prediction by soft regression. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(11): 2568-2583

[36] Cai Yijun, Li Haoxin, Hu Jian-Fang, et al. Action knowledge transfer for action prediction with partial videos//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 8118-8125

[37] Wang Xionghui, Hu Jian-Fang, Lai Jian-Huang, et al. Progressive teacher-student learning for early action prediction//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA, 2019: 3556-3565

[38] Pang Guoliang, Wang Xionghui, Hu Jian-Fang, et al. DBDNet: Learning bi-directional dynamics for early action prediction//Proceedings of the International Joint Conference on Artificial Intelligence. Macao, China, 2019: 897-903

[39] Wu Xinxiao, Wang Ruiqi, Hou Jingyi, et al. Spatial–temporal relation reasoning for action prediction in videos. International Journal of Computer Vision, 2021, 129(5): 1484-1505

[40] Tao Zhiqiang, Bai Yue, Zhao Handong, et al. Adversarial memory network for action prediction. arXiv preprint arXiv: 2112.09875, 2021

[41] Li Tianjiao, Luo Yang, Zhang Wei, et al. HARDerNet: Hardness-guide discrimination network for 3D early activity prediction. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(12): 12112-12126

[42] Alayrac Jean-Baptiste, Recasens Adrià, Schneider Rosalia, et al. Self-supervised multimodal versatile networks//Proceedings of the International Conference on Neural Information Processing Systems. Vancouver, Canada, 2020: 25-37

ZHANG Hong-Bo, Ph. D., professor. His research interests include computer vision and action understanding.

CHEN Jie, M. S. candidate. Her research interests include computer vision and early action recognition.

ZHENG Bo-Sheng, M. S. candidate. His research interests include computer vision and early action recognition.

LIU Jing-Hua, Ph. D., associate professor. Her research interests include machine learning and data mining.

DU Ji-Xiang, Ph. D., professor. His current research interests mainly include pattern recognition and machine learning.

SUN Zhen-Zhen, Ph. D., lecturer. Her research interests include machine learning and feature selection.

Background

Early Action Recognition (EAR) is an important yet challenging task in the field of video understanding. It aims to recognize human actions based on partially observed video segments, enabling earlier and faster decision-making in applications such as intelligent surveillance and human-computer interaction. Unlike traditional action recognition tasks that rely on complete video observations, EAR must predict action categories under conditions of incomplete and often ambiguous visual information.

In recent years, considerable progress has been made through the adoption of deep learning techniques, including 3D convolutional networks, recurrent models, and Transformer-based architectures. Future feature prediction has emerged as a promising direction, where models aim to reconstruct unobserved future representations from early observations to enhance prediction accuracy. Nevertheless, existing approaches primarily supervise learning by minimizing the distance between predicted and real future features, often neglecting the need to enhance the distinctiveness between predicted features and early observed features, thereby limiting the model's discriminative ability.

This study addresses the aforementioned gap by proposing a

Transformer-based framework that integrates future feature prediction with contrastive learning. By explicitly encouraging the predicted features to be simultaneously similar to the ground-truth future and dissimilar to the early observations, the model improves its representational capacity and recognition accuracy. Extensive experiments conducted on the Something-Something V2 and UCF101 datasets demonstrate that the proposed method achieves substantial improvements over existing state-of-the-art approaches, thus advancing the frontier of early action recognition research.

This work is a part of the National Natural Science Foundation of China (No. 61871196, 62306121), Natural Science Foundation of Fujian Province of China (No. 2025J01177) and the Natural Science Foundation of Xiamen of China (No. 3502Z202373040). These three foundations are trying to find new methods for understanding human action in video, including human action recognition in video surveillance, action temporal segmentation, and action quality assessment. Early action recognition method is the important technology for these tasks. Moreover, in the field of action understanding, our group published more than 20 related papers in the last two years and hold 4 related Chinese patents.