视频文本语义对齐与全视频依赖的弱监督时序动作定位

党伟超，裴丽仙，高改梅，刘春霞
（太原科技大学计算机科学与技术学院，山西 太原 030024）

摘　要：针对现有弱监督时序动作定位（WTAL）研究存在的未充分利用动作的时序特性、全局特性和动作语义一致性等问题，提出视频文本语义对齐与全视频依赖的方法（FVD-ALM），充分利用多源信息以提升动作定位的准确性和鲁棒性。首先，依托膨胀卷积扩大模型的感受野，结合注意力机制对视频内动作的变化实施精确的特征强化，确保获得准确的时序特征，捕捉动作的动态变化。随后，采用基于高斯混合模型（GMM）的期望最大化（EM）算法提取并强化视频中的全局信息，生成精确的时序激活图，理解视频的整体内容，辅助动作的定位过程。最后，设计视频文本语义对齐模块，结合动作标签中的文本信息全面理解动作，训练模型补全描述动作的文本信息，增强模型对动作类别一致性的认知并有效区分不同动作类别。实验结果表明，在 THUMOS14 和 ActivityNet1.3 这两个主流数据集上，该方法均有效，其中在 THUMOS14 上实现了 39.1%的均值平均精度（mAP），比 DTRP-Loc 方法提高了 2.0 百分点，证实了结合多源信息的方法能够显著提高动作定位的准确性，为 WTAL 任务提供了一种有效的解决方案。

关键词：弱监督动作定位；视频文本语义对齐；全视频依赖；伪标签生成；多模态融合
源代码链接：https://github.com/peilixian/FVD-ALM
中图分类号：TP391.4　　**文献标志码：**A　　**DOI：**10.19678/j.issn.1000-3428.0252238

Weakly-Supervised Temporal Action Localization via Video-Text Semantic Alignment and Full Video Dependency

DANG Weichao, PEI Lixian, GAO Gaimei, LIU Chunxia
(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, Shanxi, China)

【Abstract】In response to the challenges in existing Weakly-supervised Temporal Action Localization (WTAL) research, such as the underutilization of action temporal characteristics, global properties, and action semantic consistency, a method called Weakly-Supervised Action Localization via Video-Text Semantic alignment and Full Video Dependency (FVD-ALM) is proposed. First, dilated convolution networks are used to expand the receptive field of the model, and attention mechanisms are utilized to precisely enhance the temporal features of action instances, ensuring accurate temporal feature extraction. Subsequently, an Expectation-Maximization (EM) algorithm based on a Gaussian Mixture Model (GMM) is applied to extract and enhance global information from the video, generating accurate temporal class activation maps to aid in the action localization process. Finally, a video-text semantic alignment module is designed to comprehensively understand actions by combining textual information with action labels. The model is trained to complete textual descriptions of actions, thereby enhancing its cognitive ability for action-category consistency and effectively distinguishing different action categories. Experimental results on the THUMOS14 and ActivityNet1.3 datasets confirm the effectiveness of this method, achieving an average mean Average Precision (mAP) of 39.1% on THUMOS14, which is a 2.0 percentage points improvement over the DTRP-Loc method. This demonstrates that the method of integrating multisource information significantly improves the accuracy of action localization and provides an effective solution for weakly-supervised action localization tasks.

【Key words】Weakly-supervised Temporal Action Localization (WTAL); video-text semantic alignment; full video dependency; pseudo-label generation; multimodal fusion

0　引言

随着视频内容的爆炸性增长，高效分析和理解视频内容的技术变得至关重要。在视频动作定位领域，一个关键任务是在未剪辑视频中快速准确地定位和识别动作实例。传统的全监督方法[1-3]依赖大

基金项目：山西省自然科学基金（202203021211194，202403021221141）。
作者简介：党伟超（CCF 会员），男，副教授、博士，主研方向为智能计算、软件可靠性、视频理解；裴丽仙（通信作者），硕士研究生；高改梅，副教授、博士；刘春霞，副教授。
收稿日期：2025-03-17　　**修回日期：**2025-06-06　**E-mail：**s202320211010@stu.tyust.edu.cn

量人工标注的细粒度数据，不仅耗时耗力，还容易受到标注偏差的影响，限制了其在实际场景中的应用。因此，弱监督视频动作定位方法[4-7]应运而生，此类方法仅需视频级别的标注，大幅减少了对细粒度标签的依赖。

然而，现有的弱监督方法大多依赖于单一模态的信息，未能充分利用多模态数据中的丰富语义信息。此外，现有方法在处理视频全局依赖和跨视频信息传播方面存在不足。

首先，现有方法在动作边界的精确定位上存在不足。大多数方法采用多实例学习(MIL)框架[8-9]，忽略动作实例的完整时序跨度，导致动作边界的定位不够准确。某些动作实例可能跨越多个片段，但现有方法往往只能识别出片段级别的动作概率，无法精确捕捉动作的起始和结束时间。此外，现有方法在处理复杂动作场景时，片段的时序信息不完整导致难以有效利用动作的完整动态信息，加剧动作边界定位的不准确。针对这一问题，设计一个多层膨胀卷积网络，设定不同的膨胀率捕获不同时间尺度上的特征，并通过注意力机制动态调整不同时间尺度的重要性。这种方法能够有效扩大感受野，提高特征表示的质量，提升动作边界的精确定位能力。

其次，伪标签方法生成片段级别的伪标签来模拟全监督学习环境，但这些伪标签的质量往往不高，无法有效指导模型学习，影响动作类别的区分能力。基于片段级别特征的伪标签方法[4,8]未能充分利用视频的全局信息，导致生成的伪标签在区分背景和动作实例时存在模糊性。本文采用期望最大化(EM)算法提取视频中的代表性片段，并基于这些代表性片段生成全局上下文信息，同时借助双向随机游走技术优化特征传播，生成高质量的伪标签。这种方法能够充分利用视频的全局信息，有效区分背景和动作实例，提高动作类别的区分能力。

最后，现有方法在利用视频数据的深层次语义信息方面存在不足。视频内容不仅包含视觉信息，还蕴含丰富的语义信息，如动作的描述性文本。然而，大多数弱监督时序动作定位(WTAL)方法仅依赖于视频本身的视觉特征，未能充分利用这些多模态信息。动作标签文本中包含的动作语义信息可以为动作定位提供重要线索，但现有方法未能有效整合这些信息，导致模型对动作类别的认知不够全面。此外，现有的多模态融合方法大多平等对待不同模态的特征，未能有效利用不同模态之间的互补性，进一步限制了模型的性能。针对这一问题，设计一个视频文本语义对齐模块，间接利用生成的目标文本信息，提高 WTAL 性能。这种方法能够充分利用动作标签文本中的语义信息，增强模型对动作类别的认知，提高动作定位的准确性。

综上，本文提出一种新的 WTAL 方法 FVD-ALM。该方法从 3 个关键维度对动作定位进行优化：首先，通过时序特征增强与融合模块，借助膨胀卷积和注意力机制，扩大模型的感受野，精准捕获动作实例的长期时序依赖性，并融合 RGB 与光流特征，提升模型对时序信息的捕捉能力和特征表示的质量；其次，利用跨视频伪标签指导模块，基于高斯混合模型(GMM)和 EM 算法，提取视频的全局信息，生成高质量的伪标签，为动作定位提供精准指导，增强动作定位的准确性和鲁棒性；最后，借助视频文本语义对齐模块，补全视频描述中缺失的动作关键词，强化模型对动作相关片段的识别能力，增强模型对动作类别一致性的认知，有效区分不同动作类别。

1 相关工作

1.1 全监督时序动作定位

全监督技术依赖每一帧的时间序列标注，实现时序动作定位。受目标检测框架的启发，一些工作采用两阶段框架[9-12]解决时序动作定位问题，即首先生成动作提议，然后将这些提议输入分类模块中。另一些方法利用可训练的提议架构确定动作实例的开始和结束时间。虽然这些方法展现了卓越的性能，但它们严重依赖于精确的时序注释。

1.2 弱监督时序动作定位

由于在完全监督情况下人工标记耗时且容易出错，越来越多的研究关注 WTAL。

1.2.1 多模态融合

DTRP-Loc[13]提出辅助模态加强主导模态的具体维度，并增强两种模态的重叠部分。De-FDN[14]利用不同模态之间的相关性和互补性从视频数据中提取高质量的语义特征。W-TALC[15]、CoLA[16]和 FTCL[17]方法基于对比学习的原理，确保不同类别特征间的距离大于相同类别特征间的距离。FE-FBS[18]构建时空合作增强方案，利用残差图卷积捕捉当前片段与其他片段之间的空间和时间依赖性，生成具有空间和时间关联的片段特征。ACM-Net[19]提出了一个类不可知的动作上下文分支，用以解决动作类别间的混淆问题。CO2-Net[20]利用跨模态共识模块，借助辅助模态的信息过滤掉主模态中的信息冗余。COPL[21]、DBSMR[22]和 AFPS[23]在训练中使用额外的弱信息，获得了显著

的改进。为了克服现有方法平等对待两种模态特征的不足，本文结合了膨胀卷积扩大感受野和多模态增强特征，这种方法能有效捕捉动作实例间的时序依赖性，并增强光流特征，更好地利用时序信息。同时，借助增强的光流特征增强 RGB 特征，确保两种模态特征之间的一致性。

1.2.2 伪标签方法

ISSF[24]通过推断显著的片段特征来生成高质量的伪标签。然而，它对显著性推断模块的依赖较强，可能会影响伪标签的质量。ASM-Loc[4]利用动作提议生成的伪标签提供细粒度的监督，以改善动作边界的预测。UGCT[25]提出了一种不确定性引导的协作训练策略，通过模态协作学习和不确定性估计来生成伪标签。EM-MIL[26]则将伪标签生成放入 EM 框架中。然而，上述方法都单独处理每个片段以生成伪标签，未充分利用上下文信息，缺乏对动作的整体理解，不能有效区分模糊片段，影响了定位效果。为了解决这些问题，本文对比了 k-means、凝聚聚类、谱聚类等基于上下文信息相似度的聚类方法，选择 EM 算法挖掘每个视频中的代表性片段，并与视频记忆库中的片段相结合，有效地在视频片段间传播信息，生成更高质量的伪标签。本文利用视频的全局上下文信息，提高了动作实例的检测能力，有效提升了动作定位的准确性和鲁棒性。

1.2.3 视觉-语言模型

近年来，人们越来越关注视觉与语言如何相互作用的研究，如视觉语言预训练(VLP)[27-28]通过具有一致上下文信息的大规模图像-文本对数据集学习联合表示。CLIP(Contrastive Language-Image Pre-training)[27]作为这一领域的典型研究，将包含上下文信息的图像-文本对分别映射到视觉和文本编码器中，实现了跨模态的联合表示学习。CLIP 在图像分类、语义分割、多模态学习和视觉问题回答[29]等多个图像理解任务中取得了巨大成功。但是，它没有广泛探索如何在 WTAL 任务中充分利用封装在动作标签文本中的信息。本文设计一个新方法，探索如何利用动作标签中的文本信息来提升 WTAL 的性能。该方法间接利用生成的目标文本信息，提高 WTAL 的性能。

2 方法

在本节中，首先定义 WTAL 任务，其次介绍时序特征融合与增强策略，然后阐述跨视频上下文伪标签指导和视频文本对齐的方法，最后提供总体训练目标和推理过程的相关细节。本文提出的整体框架如图 1 所示(彩色效果见《计算机工程》官网 HTML 版，下同)，主要包含时序特征增强与融合、跨视频伪标签指导、视频文本对齐 3 个核心模块，在视频上下文提取和跨视频上下文融合部分，用虚线表示期望(E)步骤，实线表示最大化(M)步骤。模型通过膨胀卷积和注意力机制来增强融合视频特征，利用 GMM 和 EM 算法提取全局信息生成用于指导的伪标签，并结合视频文本语义对齐技术补全动作描述的文本信息，以提高动作定位的精确度。

2.1 问题定义

在 WTAL 任务中，同一个视频可能包含一个或多个动作类别。每个视频 V 可以分成一系列不重叠的片段：$V=\{V_t\}_{t=1}^{T}$，其中，T 是视频中的片段的数量，V_t 表示视频中的第 t 个片段，用它们的动作类别 $\{y_t\}_{t=1}^{T}$ 进行标注，其中 y_t 是表示视频 V_t 中存在或不存在某类动作的二进制向量。主要目的是利用弱标签训练数据集 $\{V_t,y_t\}_{t=1}^{T}$ 来计算测试视频中的一组动作实例 $\{(c,q,t_s,t_e)\}$，其中，c 表示预测类，q 表示置信度得分，t_s 和 t_e 分别表示当前动作实例的开始时间和结束时间。在时序动作定位领域中，常采用 I3D[30]、X3D[31]、UNet-3D[32]和双流卷积网络对视频进行特征提取。X3D 利用 3D 卷积捕捉视频中的空间和时间特征，而 UNet-3D 则因其结合上下文和空间信息的优势，作为提取视频特征的方法。双流卷积网络侧重于结合 RGB 流和光流建模视频的时序和空间信息。I3D 网络在传统双流网络的基础上增加了 inception 结构，使得网络能更有效地获取 3D 信息，并在分类性能和运行效率方面实现显著的提升。因此，本文模型采用 I3D 网络进行特征提取。将视频分为多个 16 帧的片段，获取 RGB 特征 $\boldsymbol{X}_r\in\mathbb{R}^{T\times D}$ 和光流特征 $\boldsymbol{X}_f\in\mathbb{R}^{T\times D}$，其中，$D$ 表示特征维度，T 为不重叠片段数。$\boldsymbol{X}_{r,n}$ 和 $\boldsymbol{X}_{f,n}$ 分别表示视频中第 n 个片段的 RGB 特征和光流特征。

2.2 时序特征增强与融合

时序特征增强和融合这两个关键步骤提升了动作与背景的区分能力，提高了动作定位的准确性。具体而言，时序特征增强部分利用多层膨胀卷积技术扩大模型的感受野，精准捕获不同时间尺度上的动作特征，时序特征融合部分则通过共享卷积层融合 RGB 特征与增强的光流特征生成注意力权重序列以强化动作相关特征，提升特征表示的一致性。

2.2.1 时序特征增强

采用多层膨胀卷积技术扩大模型的感受野，以便捕获更广泛的时空信息。设计这一模块使模型能够识别动作片段之间的时间依赖性，并据此增强特

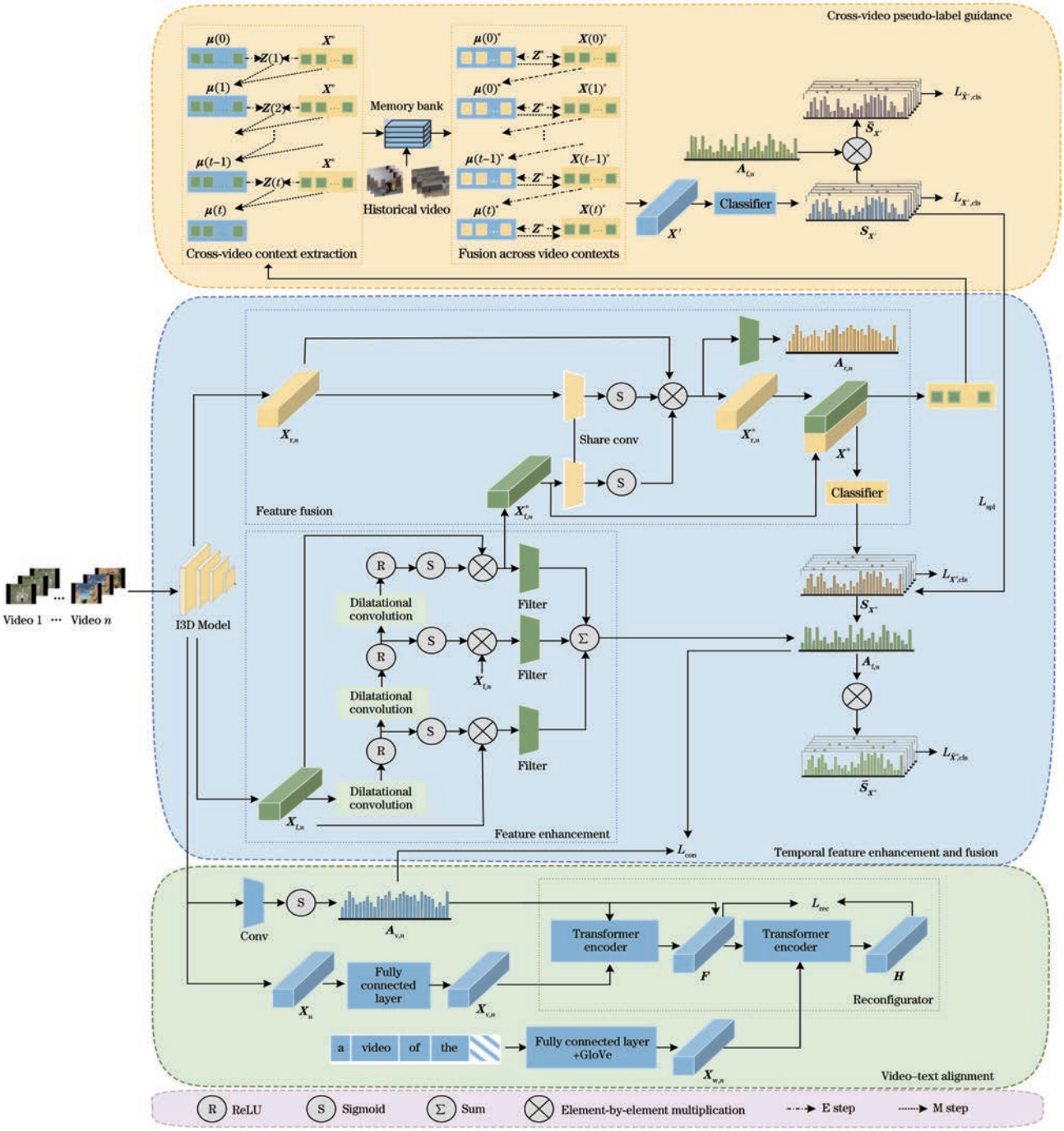

图 1 整体框架

Fig. 1 Overall framework

征表示。该模块构建一个 K 层的膨胀卷积网络，通过不同的膨胀值来捕获不同时间尺度上的特征。对于第 k 层，将输入特征 $\boldsymbol{M}_{n,k-1}\in\mathbb{R}^{D\times T}$ 通过带有扩张率 $r=2^{k-1}$ 的膨胀卷积操作 $f_{\text{dilatedConv}}$ 和 ReLU 激活函数进行处理，得到新的特征表示 $\boldsymbol{M}_{n,k}\in\mathbb{R}^{D\times T}$，其中 $\boldsymbol{M}_{n,0}$ 表示初始的光流特征 $\boldsymbol{X}_{\text{f},n}$。这个过程可以用以下公式表示：

$$\boldsymbol{M}_{n,k}=\text{ReLU}(f_{\text{dilatedConv}}(\boldsymbol{M}_{n,k-1},r=2^{k-1})),\quad k=1,\cdots,K \tag{1}$$

然后，对最后一层的输出 $\boldsymbol{M}_{n,k}$ 进行 Sigmoid 函数处理，将处理结果与原始光流特征进行逐元素乘法，得到增强的光流特征 $\boldsymbol{X}^{*}_{\text{f},n}$，表示为：

$$\boldsymbol{X}^{*}_{\text{f},n}=\sigma(\boldsymbol{M}_{n,k})\otimes\boldsymbol{X}_{\text{f},n} \tag{2}$$

式中：$\otimes$ 表示逐元素乘法；σ 表示 Sigmoid 激活函数。同时，该模块还包含一个注意力权重生成机制，对每层的输出先进行 Sigmoid 函数处理和逐元素乘法，再通过一个过滤模块生成注意力权重 $\boldsymbol{A}_{\text{f},n,k}\in\mathbb{R}^{T\times 1}$。随后对这些权重进行加权平均，计算出最终的时间注意力权重 $\boldsymbol{A}_{\text{f},n}$：

$$\boldsymbol{A}_{\text{f},n,k}=f_{\text{filter}}(\sigma(\boldsymbol{M}_{n,k})\otimes\boldsymbol{X}_{\text{f},n}),k=1,\cdots,K \tag{3}$$

$$\boldsymbol{A}_{\mathrm{f},n}=\sum_{k=1}^{K}\omega_k\boldsymbol{A}_{\mathrm{f},n,k},k=1,\cdots,K \tag{4}$$

式中：f_{filter} 表示过滤模块；ω_k 表示归一化的权重系数，确保所有系数之和为 1。过滤模块由 3 个 1D 卷积层后跟 Sigmoid 函数组成，进一步提炼和强化与动作相关的特征部分。这一过程使得模型能够综合多层网络的信息，获得更全面的时间注意力表示，为动作识别和定位提供更丰富的时序上下文信息。

2.2.2 时序特征融合

定义共享卷积层 f_{conv} 同时处理 RGB 特征 $\boldsymbol{X}_{\mathrm{r},n}$ 和增强的光流特征 $\boldsymbol{X}^{*}_{\mathrm{f},n}$，生成两个注意力权重序列。这些权重序列通过逐元素乘法的方式增强原始 RGB 特征，生成增强的 RGB 特征 $\boldsymbol{X}^{*}_{\mathrm{r},n}$。以下是该过程的数学表达：

$$W_{\mathrm{RGB}}=\sigma(f_{\mathrm{conv}}(\boldsymbol{X}_{\mathrm{r},n})) \tag{5}$$

$$W_{\mathrm{Flow}}=\sigma(f_{\mathrm{conv}}(\boldsymbol{X}^{*}_{\mathrm{f},n})) \tag{6}$$

式中：Sigmoid 激活函数将卷积层的输出转换为权重序列。然后，用这两个权重序列增强原始 RGB 特征，得到增强的 RGB 特征 $\boldsymbol{X}^{*}_{\mathrm{r},n}$：

$$\boldsymbol{X}^{*}_{\mathrm{r},n}=\boldsymbol{X}_{\mathrm{r},n}\otimes W_{\mathrm{RGB}}\otimes W_{\mathrm{Flow}} \tag{7}$$

将两个权重序列与原始 RGB 特征相融合，突出显示与动作相关的特征，抑制与背景相关的特征。接着，将增强的 RGB 特征 $\boldsymbol{X}^{*}_{\mathrm{r},n}$ 输入一个过滤模块中生成空间注意力权重 $\boldsymbol{A}_{\mathrm{r},n}$：

$$\boldsymbol{A}_{\mathrm{r},n}=f_{\mathrm{filter}}(\boldsymbol{X}^{*}_{\mathrm{r},n}) \tag{8}$$

通过这种特征融合机制，模型能够更有效地区分视频中的动作和背景，提高动作定位的准确性。

2.3 跨视频伪标签指导

提出跨视频伪标签指导分支，通过提取和融合视频的全局上下文信息，实现对视频的全局依赖，生成高质量的时序类激活序列图作为伪标签指导模型学习，使模型更全面地理解和定位视频中的动作实例。

2.3.1 跨视频上下文提取

具有代表性的片段应能描述同类中的大部分片段，因此，为了提取视频中的代表性片段，使用基于 GMM 的方法。GMM 的数学表达式为：

$$p(\boldsymbol{x}^{*}_{n})=\sum_{i=1}^{u}z_{ni}N(\boldsymbol{x}^{*}_{n}\mid\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i) \tag{9}$$

式中：$\boldsymbol{x}^{*}_{n}$ 是第 n 个片段的特征向量；u 是高斯分布的个数；$\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Sigma}_i$ 分别是第 i 个高斯分布的均值和协方差矩阵；z_{ni} 是第 n 个片段属于第 i 个高斯分布的概率。为了从视频数据中学习 GMM 的参数，采用一种迭代优化算法——EM 算法[33]。EM 算法用来捕捉每个视频片段特征的分布，并从中提取代表性片段。算法每次迭代分为两个步骤：E 步骤和 M 步骤。在 E 步骤中，给定观测数据和当前参数估计，计算潜在变量的期望值，即使用 EM 算法来估计视频特征的 GMM，将每个高斯分布的均值视为一个潜在的代表性片段。算法开始时随机初始化高斯分布的均值 $\boldsymbol{\mu}(0)\in\mathbb{R}^{N\times D}$，然后计算高斯函数第 t 次迭代的权重 $\boldsymbol{Z}(t)\in\mathbb{R}^{T\times N}$，这个权重反映每个片段属于各个高斯分布的概率分布，权重的计算公式为：

$$\boldsymbol{Z}(t)=\mathrm{Softmax}(\lambda\|\boldsymbol{X}^{*}\|_2\|(\boldsymbol{\mu}(t-1))^{\mathrm{T}}\|_2) \tag{10}$$

式中：$\boldsymbol{X}^{*}\in\mathbb{R}^{T\times 2D}$ 表示综合的视频特征，由时序特征增强与融合模块的 $\boldsymbol{X}^{*}_{\mathrm{r},n}$ 和 $\boldsymbol{X}^{*}_{\mathrm{f},n}$ 两种特征合并得到；λ 是一个控制分布平滑度的超参数；$\|\cdot\|_2$ 表示 L2 范数。接着，在 M 步骤，最大化在 E 步骤计算的期望值，更新参数估计，即使用这些权重更新高斯分布的均值 $\boldsymbol{\mu}$，公式如下：

$$\boldsymbol{\mu}(t)=\|(\boldsymbol{Z}(t))^{\mathrm{T}}\|_1\boldsymbol{X}^{*} \tag{11}$$

式中：$\|\cdot\|_1$ 表示 L1 范数。均值向量的数量远少于视频片段的数量（$T\gg N$），交替执行 E 步骤和 M 步骤是一种高效的跨视频上下文提取方法。

经过两次 EM 迭代，模型最终能够获得一组代表性片段，这些片段被保存在存储库中作为后续特征融合的基础。

2.3.2 跨视频上下文融合

在跨视频上下文融合部分，融合代表性片段的信息更新视频特征，从而提高动作定位的准确性。融合过程的关键是如何将提取出的代表性片段的特征有效地传播到当前视频的所有片段中。关键的传播过程使用亲和力 $\boldsymbol{Z}^{*}$ 进行随机游走操作（$\boldsymbol{Z}^{*}\boldsymbol{\mu}^{*}$），指导代表性片段特征融合到原始视频特征中，即存储库中给定上下文代表性片段 $\boldsymbol{\mu}^{*}$ 和原始特征 $\boldsymbol{X}^{*}$，通过加权和计算更新得到全局上下文融合特征 $\boldsymbol{X}'$，公式如下：

$$\boldsymbol{X}'=\omega\boldsymbol{Z}^{*}\boldsymbol{\mu}^{*}+(1-\omega)\boldsymbol{X}^{*} \tag{12}$$

式中：ω 是一个权衡参数，用于平衡特征传播和原始特征之间的偏离度；$\boldsymbol{Z}^{*}$ 由式（10）直接计算得到，且基于存储库中的代表片段。

尽管上下文片段与大多数同类片段有很高的相似性，但通过一次传播过程，仍无法实现上下文信息的完全融合。为了充分融合代表性片段的特征，使用双向随机游走[34]多次迭代逐步更新特征。在第 t 次迭代中，传播过程的公式可以表示为：

$$\boldsymbol{\mu}(t)^{*}=\omega\|(\boldsymbol{Z}^{*})^{\mathrm{T}}\|_1\boldsymbol{X}(t-1)^{*}+(1-\omega)\boldsymbol{\mu}(0)^{*} \tag{13}$$

$$\boldsymbol{X}(t)^{*}=\omega\boldsymbol{Z}^{*}\boldsymbol{\mu}(t)^{*}+(1-\omega)\boldsymbol{X}(0)^{*} \tag{14}$$

式中：$\boldsymbol{\mu}(0)^*$ 和 $\boldsymbol{X}(0)^*$ 分别表示初始的代表性片段和视频片段增强与融合的特征。

式(13)和式(14)可视为一个 EM 过程，它固定亲和力 $\boldsymbol{Z}^*$，交替更新 $\boldsymbol{X}^*$ 和 $\boldsymbol{\mu}^*$。因此，该方法让模型不仅能传播代表性片段的特征，还能作为促进同一类别内不同特征之间传播的桥梁。反复执行这个过程，确保代表性片段的特征能充分融合以得到更高质量的全局上下文融合特征。

2.3.3 伪标签指导

将时序增强与融合特征 $\boldsymbol{X}^*$ 和全局上下文融合特征 $\boldsymbol{X}'$ 这两种特征在 MIL 框架下投影到分类头中生成对应的时序类别激活序列 $\boldsymbol{S}_{\boldsymbol{X}^*}$ 和 $\boldsymbol{S}_{\boldsymbol{X}'}$：

$$\boldsymbol{S}_{\boldsymbol{X}^*}=f_{\mathrm{cls}}(\boldsymbol{X}^*) \tag{15}$$

$$\boldsymbol{S}_{\boldsymbol{X}'}=f_{\mathrm{cls}}(\boldsymbol{X}') \tag{16}$$

此外，根据背景抑制的方法，应用注意力权重 $\boldsymbol{A}_{\mathrm{f},n}$ 抑制背景片段对动作片段的响应。背景抑制的结果 $\bar{\boldsymbol{S}}_{\boldsymbol{X}^*}$ 可以通过 $\bar{\boldsymbol{S}}_{\boldsymbol{X}^*}=\boldsymbol{A}_{\mathrm{f},n}\otimes\boldsymbol{S}_{\boldsymbol{X}^*}$ 得到。通过同样方式得到背景抑制的 $\bar{\boldsymbol{S}}_{\boldsymbol{X}'}$。将全局上下文融合的时序类别激活序列 $\boldsymbol{S}_{\boldsymbol{X}'}$ 作为细粒度的伪标签，来指导和优化 $\boldsymbol{S}_{\boldsymbol{X}^*}$。

2.4 视频文本对齐

通过补全视频描述中缺失的动作关键词，强化模型对动作相关片段的识别能力，引入自监督一致性约束确保模型在不同目标下对动作区域的一致关注，从而提高动作定位的准确性和鲁棒性。

2.4.1 视频与文本嵌入

首先将视频特征 $\boldsymbol{X}_n\in\mathbb{R}^{T\times 2D}$ 通过全连接层转化为嵌入特征 $\boldsymbol{X}_{\mathrm{v},n}\in\mathbb{R}^{T\times M}$，其中 $M=512$ 为嵌入后的特征维度且 $\boldsymbol{X}_{\mathrm{v},n}=\mathrm{emb}(\boldsymbol{X}_n)$。然后利用由多个卷积层和 Sigmoid 函数组成的注意力机制，从视频特征中筛选出与动作文本相关的视频片段，并得到注意力权重 $\boldsymbol{A}_{\mathrm{v},n}\in\mathbb{R}^{T\times 1}$。此外，构造一个基于动作标签的句子并隐藏句子中的关键词汇，通过 GloVe[35] 嵌入和全连接层将这些带有遮蔽词汇的句子转化为语句特征嵌入 $\boldsymbol{X}_{\mathrm{w},n}\in\mathbb{R}^{L\times M}$，其中 L 是句子的长度。随后，利用 Transformer 重构器预测这些遮蔽的词汇[36]，完成句子的补全。这一过程如图 2 所示。

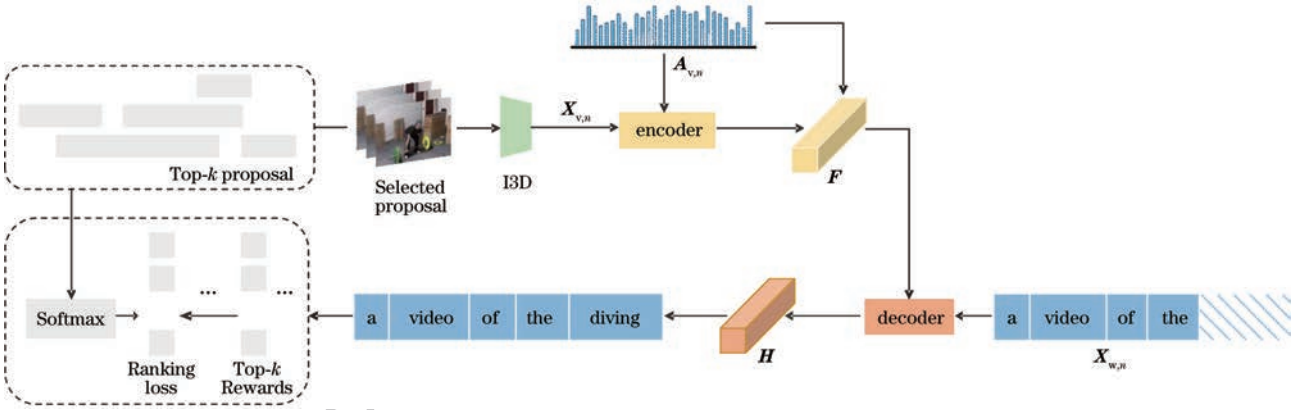

图 2 Transformer 重构器

Fig. 2 Transformer reconstructor

遮蔽词汇重构任务的核心在于通过隐藏句子中的关键词汇，迫使模型从视频特征中推断出这些缺失的词汇。这些关键词汇通常是动作的核心描述词，如“跳高”或“标枪投掷”。通过这种方式，模型不仅需要理解视频中的视觉内容，还需要将其与文本描述中的语义信息进行对齐。这一过程显著增强了模型对动作类别的认知能力，使其能够更准确地识别和定位视频中的动作实例。例如，提示模板是“一个人在跳高”，模型需要从视频中识别出“跳高”这一动作的特征，从而更准确地定位该动作的起始和结束时间。通过这种自监督学习的方式，模型能够学习到更丰富的语义特征，这些特征有助于区分不同动作类别，从而提高动作定位的准确性。此外，遮蔽词汇重构任务还提供了一种额外的监督信号，使得模型在训练过程中能够更好地捕捉视频内容与文本描述之间的一致性，进一步提升模型的鲁棒性和泛化能力。

这一过程中计算前景视频特征 $\boldsymbol{F}$ 和多模态 $\boldsymbol{H}$ 数学表达式如下：

$$\boldsymbol{F}=\mathrm{Encoder}(\boldsymbol{X}_{\mathrm{v},n},\boldsymbol{A}_{\mathrm{v},n})=\sigma\left(\frac{\boldsymbol{X}_{\mathrm{v},n}\boldsymbol{W}_{\mathrm{q}}(\boldsymbol{X}_{\mathrm{v},n}\boldsymbol{W}_{\mathrm{k}})^{\mathrm{T}}}{\sqrt{D_{\mathrm{v},n}}}\boldsymbol{A}_{\mathrm{v},n}\right)\boldsymbol{X}_{\mathrm{v},n}\boldsymbol{W}_{\mathrm{v}} \tag{17}$$

$$\boldsymbol{H}=\mathrm{Decoder}(\boldsymbol{X}_{\mathrm{w},n},\boldsymbol{F},\boldsymbol{A}_{\mathrm{v},n})=\sigma\left(\frac{\boldsymbol{X}_{\mathrm{w},n}\boldsymbol{W}_{\mathrm{qd}}(\boldsymbol{F}\boldsymbol{W}_{\mathrm{kd}})^{\mathrm{T}}}{\sqrt{D_{\mathrm{v},n}}}\boldsymbol{A}_{\mathrm{v},n}\right)\boldsymbol{F}\boldsymbol{W}_{\mathrm{vd}} \tag{18}$$

式中：Encoder 和 Decoder 分别为 Transformer 模型中的编码器和解码器部分；$\boldsymbol{W}_{\mathrm{q}}$，$\boldsymbol{W}_{\mathrm{k}}$，$\boldsymbol{W}_{\mathrm{v}}$，$\boldsymbol{W}_{\mathrm{qd}}$，$\boldsymbol{W}_{\mathrm{kd}}$，

$\boldsymbol{W}_{\mathrm{vd}}$ 都是可学习的参数；$D_{v,n}$ 表示 $\boldsymbol{X}_{v,n}$ 的特征维度，$D_{v,n}=512$。

最终，为了实现更准确的文本补全描述，视频文本嵌入的重构损失函数可以公式化为：

$$L_{\mathrm{rec}}=\mathrm{MSE}(\boldsymbol{F},\boldsymbol{H}) \tag{19}$$

2.4.2 自监督一致性约束

通过最小化时序特征增强与融合模块和视频文本对齐模块中注意力权重之间的均方误差（MSE）损失，避免对视频中显著片段的过度依赖，提高对动作区域的整体识别能力，如式(20)所示：

$$L_{\mathrm{con}}=\mathrm{MSE}(\boldsymbol{A}_{\mathrm{f},n},\mathrm{Clip}(\boldsymbol{A}_{\mathrm{f},n}))+\mathrm{MSE}(\boldsymbol{A}_{\mathrm{v},n},\mathrm{Clip}(\boldsymbol{A}_{\mathrm{v},n})) \tag{20}$$

式中：Clip 函数表示梯度限制，用来避免在优化过程中出现梯度爆炸问题。通过这种方式，能够全面地考虑视频中与动作相关的所有片段，而不只是考虑突出的部分，从而提高动作定位的准确性和鲁棒性。这种结合视频分析和文本数据的方法为应对 WTAL 的挑战提供了新的视角。

2.5 训练目标和推理过程

设计一个损失函数来优化动作定位模型，该函数整合多实例动作分类损失、背景抑制损失、归一化损失、伪标签监督损失、注意力引导损失及一致性约束损失，以提升模型在弱监督环境下的性能。在构建损失函数时，首先引入多实例动作分类损失，它作用在模型的时序特征增强与融合和跨视频伪标签指导分支，帮助模型区分不同的动作类别。这一损失可以表示为：

$$L_{\mathrm{cls}}=L_{\boldsymbol{X}^{*},\mathrm{cls}}+L_{\boldsymbol{X}',\mathrm{cls}} \tag{21}$$

用不同的损失项来平衡对模型训练的贡献，$L_{\boldsymbol{X}^{*},\mathrm{cls}}$ 和 $L_{\boldsymbol{X}',\mathrm{cls}}$ 分别对应时序特征增强与融合的动作分类损失和跨视频伪标签指导分支的动作分类损失。

接着，引入背景抑制的损失 $L_{\mathrm{bg,cls}}=L_{\bar{\boldsymbol{X}}^{*},\mathrm{cls}}+L_{\bar{\boldsymbol{X}}',\mathrm{cls}}$，目的是降低模型对背景片段的误分类率，增强模型对动作的关注。这一损失项与多实例动作分类损失相结合，共同促进模型更准确识别动作类别。

为了进一步提升模型的泛化能力，引入归一化损失 $L_{\mathrm{norm}}=\dfrac{(\|\boldsymbol{A}_{\mathrm{f},n}\|_1+\|\boldsymbol{A}_{\mathrm{r},n}\|_1)}{2}$，确保模型输出的稳定性。随后，将通过跨视频伪标签指导分支生成的时序类激活序列作为伪标签，设计伪标签监督损失：

$$L_{\mathrm{spl}}=-\sum_{n=1}^{T}\log_a S_n(\boldsymbol{X}') \tag{22}$$

式中：$S_n(\boldsymbol{X}')$表示第 n 个片段的时序类激活序列与全局上下文特征增强特征之间的匹配度。此外，为了优化模型中的注意力权重，引入注意力引导损失：

$$L_{\mathrm{guide}}=\sum_{n=1}^{T}[|1-\boldsymbol{A}_{\mathrm{f},n}-S_{\mathrm{bg},n}|+|1-\boldsymbol{A}_{\mathrm{r},n}-S_{\mathrm{bg},n}|] \tag{23}$$

式中：$\boldsymbol{A}_{\mathrm{f},n}$ 和 $\boldsymbol{A}_{\mathrm{r},n}$ 分别表示第 n 个片段的光流注意力权重和 RGB 注意力权重；$S_{\mathrm{bg},n}$ 是相应的背景抑制的类激活序列。将所有这些损失函数整合，形成一个总损失函数：

$$L=L_{\mathrm{cls}}+L_{\mathrm{bg,cls}}+\lambda_1 L_{\mathrm{norm}}+L_{\mathrm{spl}}+\lambda_2 L_{\mathrm{guide}}+\lambda_3(L_{\mathrm{con}}+L_{\mathrm{rec}}) \tag{24}$$

式中：λ_1、λ_2、λ_3 都是超参数，它们综合上述所有组件，指导模型的训练。这个总损失函数不仅考虑模型在各个子任务上的表现，还平衡它们之间的权重，使得模型在整体训练过程中达到最佳训练效果。

3 实验结果与分析

本节首先展示所提方法在两个 WTAL 数据集 THUMOS14[37] 和 ActivityNet1.3[38] 上的测试结果，随后通过消融实验证明其有效性。

3.1 数据集

在两个主流的动作定位数据集 THUMOS14 和 ActivityNet1.3 上进行实验，并采用视频级别的标签来训练模型。

THUMOS14 数据集涵盖了 20 种动作类型，其验证集有 200 个未经剪辑的视频，测试集则包含 213 个未经剪辑的视频。这些视频的长度差异很大，短至几秒，长至超过一小时。每个视频可能包含多个动作实例，有超过 70%的帧为上下文帧或背景帧。选取验证集视频用于模型训练，测试集视频用于测试模型性能。

ActivityNet1.3 数据集相较于 THUMOS14 数据集，规模更为庞大，涵盖了更多与人类日常生活相关的动作，视频数量多、类别丰富，包含 200 种不同类别的动作，其中有 10 024 个未修剪的视频用于模型的训练，4 296 个未修剪的视频用于模型的性能测试，约有 36%的帧为上下文帧或背景帧，大部分视频时长在 5～10 min，50%的视频的分辨率为 1 280×720 像素，大部分视频帧率为 30 帧/s，类别主要分为个人护理、饮食、家庭活动、关怀和帮助、工作、社交娱乐、运动锻炼七大类。

3.2 评估指标

遵循标准的评估协议，通过计算不同交并比（IoU，r_{IoU}）阈值下的均值平均精度（mAP，P_{mAP}）来评

估 WTAL 的性能，即 mAP@IoU。对于 THUMOS14，设置 IoU 阈值为 0.1∶0.1∶0.7，ActivityNet1.3 的 IoU 阈值为 0.50∶0.05∶0.95。

通过计算模型在多个不同 IoU 阈值下的 mAP 来衡量其在动作定位任务上的表现。IoU 是衡量预测动作区域与实际动作区域之间重叠程度的指标，即预测区域与真实区域的交集与并集的比例，公式表示为：

$$r_{\mathrm{IoU}}(I_{\mathrm{P}},I_{\mathrm{g}})=\frac{I_{\mathrm{P}}\cap I_{\mathrm{g}}}{I_{\mathrm{P}}\cup I_{\mathrm{g}}} \tag{25}$$

式中：I_{P} 为预测动作提议；I_{g} 为真实动作提议。设定不同的 IoU 阈值，当预测动作提议和真实动作提议之间的 IoU 大于等于阈值时，计算此阈值下的平均 mAP。mAP 的公式为：

$$P_{\mathrm{mAP}}=\frac{\sum_{i=1}^{C}P_{\mathrm{AP}i}}{C} \tag{26}$$

式中：C 为动作类别总数；$P_{\mathrm{AP}i}$ 表示第 i 类动作的检测精度(AP)均值。AP 计算方式为：

$$P_{\mathrm{AP}}=\int_0^t P(R)\,\mathrm{d}R \tag{27}$$

式中：P 表示准确率；R 为召回率。准确率和召回率具体定义为：

$$\begin{cases}P=\dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}}\\[2ex] R=\dfrac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}}\end{cases} \tag{28}$$

式中：N_{TP} 表示实际动作帧被检测为动作帧的数量；N_{FP} 表示实际背景帧被检测为动作帧的数量；N_{FN} 表示实际动作帧被检测为背景帧的数量。通过计算每个动作类别 AP 的平均值，并对所有动作类别的 AP 求平均值，即可得到 mAP。

3.3 实验细节

1)实验环境。

基于视频文本语义对齐与全视频依赖的 WTAL 方法部署在 PyTorch 环境下，实验所采用的硬件及软件配置如表 1 所示。

表 1 硬件及软件配置

Table 1 Hardware and software configuration

Hardware/Software	Configuration parameter
CPU	Intel® Core™ i7-14650HX
GPU	NVIDIA GeForce RTX 4060 Laptop GPU
PyTorch	PyTorch 2.0.0
Python	Python 3.8
CUDA	CUDA 11.8
OS	Windows 10

2)特征提取。

使用的 I3D 特征提取是一种深度学习方法，它通过预训练的三维卷积神经网络从视频中提取时空特征，将 RGB 帧和光流帧输入网络并输出包含丰富时空信息的高维特征向量。在操作过程中，首先将视频采样成 RGB 帧，然后利用 TV-L1 等光流算法计算出相邻帧间的运动信息，形成光流图。

将 RGB 帧和光流图分成 16 帧一段的片段，输入 I3D 网络中，网络通过多层三维卷积层提取并融合空间和时间特征，为每个片段生成一个 1 024 维的特征向量，这些特征向量可以用于动作识别、视频分类和动作定位等多种视频理解任务。

3)THUMOS14 和 ActivityNet1.3 数据集上的实验。

使用 Adam 优化器进行模型训练，膨胀卷积中扩张率对应的层数 $k=3$。对于 THUMOS14，学习率为 0.000 5，权重衰减为 0.001，优化模型约 5 000 次迭代。对于 ActivityNet1.3，学习率为 0.000 03，优化模型约 50 000 次迭代。在时序特征增强与融合模块中，将超参数 ω_k 设为 1/3 且 $k=1$，2，3，跨视频伪标签指导中的超参数 λ，ω 分别设置为 5.0，0.5。此外，对于超参数 λ_1，λ_2，λ_3，在 THUMOS14 上分别设置为 1.0，1.0，1.5，在 ActivityNet1.3 上分别设置为 1.0，1.0，0.25。

3.4 实验结果比较与分析

在 THUMOS14 数据集上将所提出的方法与 BSN[39]、GTAN[40]、BUMR[41]、TCANet[10]、CoLA[16]、CO2-Net[20]、ASM-Loc[4]、DELU[42]、DDG-Net[43]、P-MIL[44]、DBSMR[22]、FE-FBS[18]、COPL[21]、ISSF[24]、De-FDN[14]、DTRP-Loc[13]、CCKEE[45]进行比较，在 ActivityNet1.3 数据集上与 UGCT[25]、FAC-Net[46]、DCC[47]、ASM-Loc[4]、P-MIL[44]、LPR[5]、CCKEE[45]、CFIL[48]、SRHN[49]、AFPS[23]进行比较，结果分别如表 2 和表 3 所示(最优结果加粗标示，次优结果下划线标示，下同)。

从两个数据集的实验结果来看，所提出的方法在 WTAL 领域具有显著的创新性和有效性。相比于仅使用卷积层作为分类器的基线方法，它通过结合多源信息的方法，能够更好地利用有限的标注信息挖掘出视频中动作的特征，从而提高动作定位的准确性和完整性。同时，该方法不仅适用于特定的数据集，而且能够适应不同的数据分布和标注情况，具有广泛的适用性。在 THUMOS14 数据集上，该方法的性能明显优于其他 WTAL 方法。在 mAP@0.3∶0.7方面，DTRP-Loc方法主要利用局

表 2 不同方法在 THUMOS14 数据集上的检测结果

Table 2 Detection results of different methods on the THUMOS14 dataset %

Supervision	Method	mAP@0.1	mAP@0.2	mAP@0.3	mAP@0.4	mAP@0.5	mAP@0.6	mAP@0.7	mAP@0.3:0.7
Full	BSN[39]	—	—	53.5	45.0	36.9	28.4	20.0	36.8
	GTAN[40]	69.1	63.7	57.8	47.2	38.8	—	—	—
	BUMR[41]	—	—	53.9	50.7	45.4	38.0	28.5	43.3
	TCANet[10]	—	—	60.6	53.2	44.6	36.8	26.7	44.4
Weak	CoLA[16]	66.2	59.5	51.5	41.9	32.2	22.0	13.1	32.1
	CO2-Net[20]	70.1	63.6	54.5	45.7	38.3	26.4	13.4	35.6
	ASM-Loc[4]	71.2	65.5	57.1	46.8	36.6	25.2	13.4	35.8
	DELU[42]	71.5	66.2	56.5	47.7	40.5	27.2	**15.3**	35.9
	DDG-Net[43]	72.5	67.7	58.2	49.0	41.4	27.6	14.8	38.2
	P-MIL[44]	71.8	67.5	58.9	49.0	40.0	27.1	15.1	38.0
	DBSMR[22]	69.1	62.3	53.2	44.1	34.3	—	13.7	—
	FE-FBS[18]	66.7	—	51.7	—	32.3	—	12.5	—
	COPL[21]	65.4	60.5	52.6	44.0	33.9	22.8	12.8	33.2
	ISSF[24]	72.4	66.9	58.4	49.7	41.8	25.5	12.8	37.6
	De-FDN[14]	72.4	65.9	56.8	45.8	34.8	24.1	14.0	35.1
	DTRP-Loc[13]	71.4	66.0	58.0	49.3	41.4	28.0	15.0	38.3
	CCKEE[45]	72.6	67.1	59.5	49.3	39.4	26.5	13.4	37.6
	FVD-ALM	**72.9**	**67.9**	**59.6**	**50.1**	**42.3**	**28.9**	14.6	**39.1**

表 3 不同方法在 ActivityNet1.3 数据集上的检测结果

Table 3 Detection results of different methods on the ActivityNet1.3 dataset %

Method	mAP@0.5	mAP@0.75	mAP@0.95	mAP@0.5:0.95
UGCT[25]	39.1	22.4	5.8	23.8
FAC-Net[46]	37.6	24.2	6.0	24.0
DCC[47]	38.8	24.2	5.7	24.3
ASM-Loc[4]	41.0	24.9	6.2	25.1
P-MIL[44]	41.8	25.4	5.2	25.5
LPR[5]	41.4	25.3	6.2	25.4
CCKEE[45]	41.6	25.1	**6.5**	25.3
CFIL[48]	41.8	25.8	6.2	25.7
SRHN[49]	41.7	26.1	6.1	26.2
AFPS[23]	43.9	**27.1**	6.3	27.3
FVD-ALM	**44.0**	**27.1**	**6.5**	**27.4**

部特征进行动作定位，而 FVD-ALM 方法通过整合全局信息，有助于模型理解视频的整体内容，实现 0.8 百分点的提升，这一提升幅度在该领域中具有重要意义，表明该方法在动作定位的准确性和完整性方面都有较好的表现。FVD-ALM 甚至超过一些完全监督的方法，这说明在弱监督条件下，通过该方法能够充分利用有限的标注信息，挖掘出视频中动作的有效特征，从而达到甚至超越完全监督方法的效果，这对于降低标注成本、提高动作定位效率具有重要的实际意义。而在 IoU 为 0.7 时，DELU 方法（15.3% mAP）略优于本方法（14.6% mAP）。这主要归因于 DELU 采用的不确定性建模机制能有效处理背景噪声对视频级预测的干扰。在 THUMOS14 这类动作实例较少且背景噪声复杂的数据集上，DELU 在高 IoU 阈值下的动作定位表现更突出。然而，在动作类别更丰富、背景噪声较少的 ActivityNet1.3 数据集上，本文方法展现出显著优势：在 IoU 为 0.5 和 IoU 为 0.75 时分别取得 44.0%和 27.1%的 mAP，表明本文方法更适用于动作实例密集的场景。同时，本文方法以 27.4%的平均 mAP 超越所有对比方法，在更具挑战性的 ActivityNet1.3 数据集上取得领先性能，充分验证了其优越的定位准确性和跨数据集的泛化能力。

3.5 消融实验

消融实验均在 THUMOS14 数据集上验证本文方法的有效性。

3.5.1 每个组件的有效性

框架主要包含 3 个组成部分：

1）时序特征增强与融合模块通过扩大感受野和增强光流特征来有效捕捉动作实例的时序依赖性；

2）使用跨视频伪标签指导模型学习，表示为 L_{spl}；

3）视频文本对齐以自监督的方式约束 WTAL 模型，表示为 L_{con}。

为了验证框架中每个组件的有效性，进行消融实验分析各个组件，结果如表 4 所示。具体在 4 个变体上做实验：

1）Baseline：仅使用卷积层作为分类器，而不是时序特征增强与融合模块；

2）Baseline＋L_{spl}：使用跨视频伪标签指导模型学习；

3）Baseline＋L_{spl}＋L_{con}：以自监督的方式约束模型；

4）Temporal feature enhancement and fusion＋L_{spl}＋L_{con}：最终框架，在“Baseline＋L_{spl}＋L_{con}”的基础上用时序特征增强与融合模块代替基线 WTAL 模型。

表 4 每个组件在 THUMOS14 数据集上的有效性结果

Table 4 Effectiveness of each component on the THUMOS14 dataset %

Component	mAP@0.3	mAP@0.5	mAP@0.7	mAP@0.3:0.7
Baseline	53.0	38.0	12.4	34.9
Baseline＋L_{spl}	55.4	39.8	13.4	36.7
Baseline＋L_{spl}＋L_{con}	56.0	41.0	13.9	36.9
Temporal feature enhancement and fusion＋L_{spl}＋L_{con}	59.6	42.3	14.6	39.1

通过比较“Temporal feature enhancement and fusion＋L_{spl}＋L_{con}”和“Baseline＋L_{spl}＋L_{con}”两种方法的性能，可以得出结论：使用时序特征增强与融合比一般的 WTAL 模型有更好的性能，在 THUMOS14 数据集上 mAP 提高了 2.2 百分点。逐步去除约束损失 L_{con} 和跨视频伪标签指导的 L_{spl}，所有实验在这种设置下的都逐渐下降。通过比较“Baseline＋L_{spl}”和“Baseline”两种方法，可以得出结论：提出的跨视频伪标签指导可以通过提取和融合视频的全局上下文信息有效指导 WTAL 模型学习，这在 THUMOS14 数据集上实现了 1.8 百分点的 mAP 性能提升。此外，通过对“Baseline＋L_{spl}＋L_{con}”和“Baseline＋L_{spl}”两种方法的比较，也验证一致性约束损失在视频文本对齐中的有效性。

3.5.2 不同类型的特征融合

为了验证特征增强与融合模块的有效性，比较不同类型的特征融合。不同类型特征融合的性能比较如表 5 所示。“Original RGB”表示直接输出没有任何增强的初始 RGB 特征 $\boldsymbol{X}_{r,n}$。“RGB-only”表示用 RGB 自关注权重来增强 RGB 特征，即 $\boldsymbol{X}^{*}_{r,n}=\boldsymbol{X}_{r,n}\otimes\sigma(f_{conv}(\boldsymbol{X}_{r,n}))$。“Flow-only”表示仅使用增强的光流特征来增强 RGB 特征，即 $\boldsymbol{X}^{*}_{r,n}=\boldsymbol{X}_{r,n}\otimes\sigma(f_{conv}(\boldsymbol{X}^{*}_{f,n}))$。“Non-shared”表示在 $\boldsymbol{X}_{r,n}$ 上使用卷积层 f_{conv1}，在 $\boldsymbol{X}_{f,n}$ 上使用卷积层 f_{conv2}，这两个卷积层不共享参数。“Shared”表示 $\boldsymbol{X}_{r,n}$ 和 $\boldsymbol{X}_{f,n}$ 上使用同样的卷积层 f_{conv}。

表 5 不同类型的特征融合方法在 THUMOS14 数据集上的结果

Table 5 Results of different types of feature fusion methods on the THUMOS14 dataset %

Feature fusion method	mAP@0.3	mAP@0.5	mAP@0.7	mAP@0.3:0.7
Original RGB	55.9	39.2	14.1	36.4
RGB-only	57.6	40.7	14.2	37.8
Flow-only	58.7	40.4	14.3	38.0
Non-shared	58.9	41.3	14.0	38.2
Shared	59.6	42.3	14.6	39.1

3.5.3 不同代表性片段生成策略

验证不同代表性片段生成策略的性能，结果如表 6 所示。一些传统的聚类方法，如 k-means、凝聚聚类和谱聚类与采用区分性片段生成策略的基线模型相比，检测性能取得明显的提高。模型采用 EM 代表性片段生成策略，学到了动作分类在全局视频中的代表性片段，在平均 mAP 方面实现了 39.1% 的高性能，远远优于其他聚类方法。

表 6 不同代表性片段生成策略在 THUMOS14 数据集上的结果

Table 6 Results of different representative fragment generation strategies on the THUMOS14 dataset %

Representative snippets generation strategy	mAP@0.3	mAP@0.5	mAP@0.7	mAP@0.3:0.7
Discriminative snippets	51.6	33.5	9.3	31.9
Agglomerative clustering	56.5	37.5	13.6	35.7
Spectral clustering	57.0	36.8	13.7	35.9
k-means	58.2	39.2	14.2	37.9
EM	59.6	42.3	14.6	39.1

3.5.4 不同类型语言重构器的视频文本对齐

为了验证视频文本对齐的有效性，比较不同类型的语言重构器对定位结果的影响。在表 7 中比较 3 种不同的重构器（Transformer、GRU[50] 和 LSTM[51]）可以得出结论，无论使用哪种语言重构器，视频文本对齐通过对特征增强与融合施加自监督约束提高模型性能。此外，还可以看出，Transformer 结构更适合用作该模型的语言重构器。

表 7 不同类型语言重构器在 THUMOS14 数据集上的结果

Table 7 Results of different types of language reconstructors on the THUMOS14 dataset %

Reconstructor	mAP@0.3:0.7
GRU	37.8
LSTM	38.4
Transformer	39.1

3.5.5 不同提示模板生成动作描述

比较表 8 中不同提示模板生成的动作描述对性能的影响。“a video of action [CLS]”提供了一个明确的语义提示，告诉模型这是一个关于动作的视频描述，适合于动作的初步识别和定位。“a [CLS]”更加简洁，直接聚焦于动作本身，让模型更专注于核心动作特征，但由于缺乏足够的语义背景，模型对动作细节的区分能力稍弱。“a video of the [CLS]”则提供了一个更具体的语义背景，让模型更关注动作的具体细节。这些结果表明，使用具有更丰富语义信息的视频文本对齐来约束 WTAL 模型是必要的。

表 8 不同提示模板生成动作描述在 THUMOS14 数据集上的结果

Table 8 Results of different prompt templates generating action descriptions on the THUMOS14 dataset %

Prompt template	mAP@0.3	mAP@0.5	mAP@0.7	mAP@0.3:0.7
a video of action [CLS]	58.7	41.5	14.4	38.6
a [CLS]	59.3	41.8	14.0	38.7
a video of the [CLS]	59.6	42.3	14.6	39.1

3.5.6 不同损失函数

为了评估不同损失函数对模型性能的影响，在 THUMOS14 数据集上逐步移除或替换损失函数，对比实验结果如表 9 所示，其中，√表示使用，×表示不使用。从表 9 可以看出，随着更多损失函数的加入，模型的性能逐步提升。通过一系列消融实验评估不同损失函数对模型性能的影响，验证每个损失函数在 WTAL 任务中的有效性和必要性，同时评估它们的协同作用。多实例动作分类损失 L_{cls} 作为基础损失函数，用于区分不同动作类别，背景抑制损失 $L_{bg,cls}$ 旨在减少模型对背景片段的误分类，增强对动作片段的关注。归一化损失 L_{norm} 用于确保模型输出的稳定性，防止过拟合和梯度爆炸，实验 1 和实验 2 的结果验证其在提升模型泛化能力方面的效果。伪标签监督损失 L_{spl} 利用生成的高质量伪标签提供细粒度的监督信息，帮助模型更准确地定位动作，通过对比实验 1 和实验 3，可以验证其在提升动作定位精度方面的有效性。注意力引导损失 L_{guide} 帮助模型聚焦于重要的视频片段，避免对不相关片段的过度关注，通过对比实验 1 和实验 4，表明其在提升动作区域识别能力方面有效果。一致性约束损失 L_{con} 通过最小化时序特征增强与融合模块和视频文本对齐模块中注意力权重之间的差异，确保模型对动作区域的一致关注，重构损失 L_{rec} 确保获得更准确的文本补全描述，对比实验 1 和实验 5 的结果，验证其在提升模型鲁棒性方面的有效性。同时，通过组合不同损失函数的实验（实验 6～16），评估损失函数的协同作用对模型性能的影响。这些消融实验不仅有助于理解不同损失函数的独立作用，还能揭示它们的协同效果，为模型的优化和改进提供依据。最终，所有损失函数组合（实验 16）的平均 mAP 达到 39.1%，这表明这些损失函数共同作用时能够最大化提升模型的性能。

表 9 不同损失函数在 THUMOS14 数据集上的结果

Table 9 Results of different loss functions on the THUMOS14 dataset %

Experiment	L_{cls}	$L_{bg,cls}$	L_{norm}	L_{spl}	L_{guide}	$L_{con}+L_{rec}$	mAP@0.3:0.7
1	√	√	×	×	×	×	22.4
2	√	√	√	×	×	×	29.6
3	√	√	×	√	×	×	31.2
4	√	√	×	×	√	×	30.3
5	√	√	×	×	×	√	31.7
6	√	√	√	√	×	×	34.5
7	√	√	√	×	√	×	33.8
8	√	√	√	×	×	√	34.9
9	√	√	×	√	√	×	35.7
10	√	√	×	√	×	√	36.9
11	√	√	×	×	√	√	36.6
12	√	√	√	√	√	×	37.6
13	√	√	√	√	×	√	38.1
14	√	√	√	×	√	√	37.8
15	√	√	×	√	√	√	38.4
16	√	√	√	√	√	√	39.1

3.6 可视化分析

图 3 展示了在 THUMOS14 数据集上使用 FVD-ALM 方法的几个示例结果。第 1 行显示了数据集中的部分样本帧，接下来的 3 行分别展示了真实动作定位结果、基线方法的结果以及提出的 FVD-ALM 方法的结果。从图中可以看出，FVD-ALM 方法的结果与真实动作片段的重叠度明显高于基线方法，能够更准确地定位动作的起始和结束时间。这表明，通过利用多源信息，FVD-ALM 方法能够更完整、更准确地捕捉动作片段，从而优于基线方法。

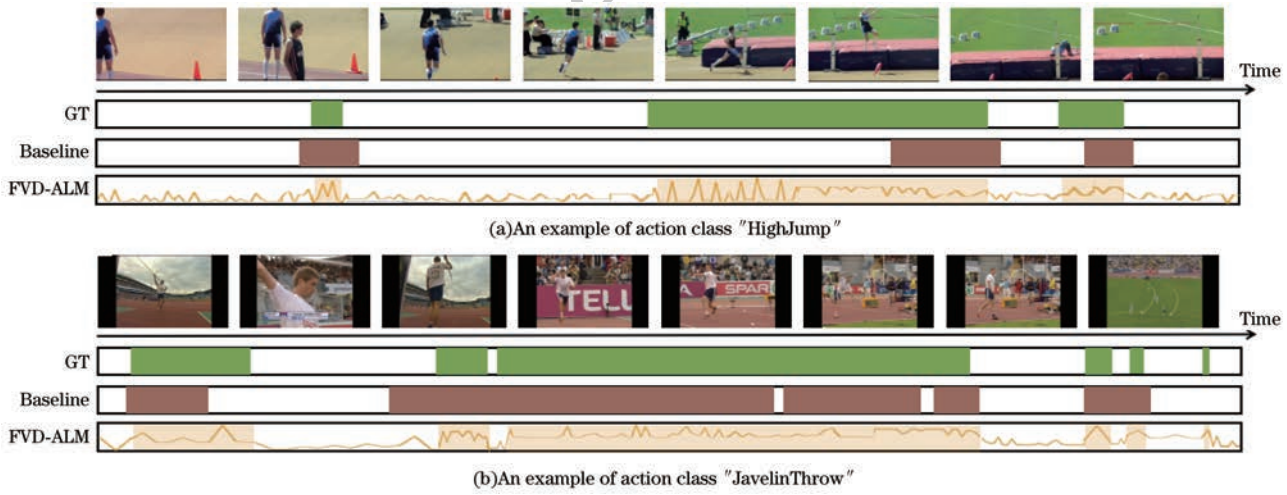

图 3 THUMOS14 数据集上的定性结果可视化

Fig.3 Visualization of qualitative results on the THUMOS14 dataset

4 结束语

针对弱监督视频动作定位问题，本文提出一种新的方法，该方法不仅依赖视频内容本身，还融合视频文本语义对齐技术，充分利用动作类别标签中的丰富语义信息，实现对视频的全面依赖。通过卷积层和注意力机制，精确强化视频中动作变化的特征，确保模型能够获得准确的动作时序特征。此外，基于 GMM 的 EM 算法帮助提取并强化视频中的全局信息，生成精确的时序激活图，辅助动作的定位。视频文本语义对齐技术的引入，进一步增强模型对动作类别一致性的认知，有效区分不同动作类别。该方法在 THUMOS14 和 ActivityNet1.3 这两个数据集上均取得较好的性能，验证其在弱监督环境下对动作定位任务的有效性。这表明融合多模态信息能显著提高动作定位的准确性，尤其是在视频内

容理解和语义信息挖掘方面。本文利用对比学习框架融合多模态信息，有效提升了定位精度。但所提方法在背景噪声较多的动作定位任务中仍有改进空间，动作跨视频依赖信息和多模态表示的融合机制，也是值得进一步探索的方向。未来，考虑优化噪声抑制策略，以提升模型在高 IoU 阈值下的鲁棒性。此外，还考虑将多模态表示 H 与增强后的特征 $\boldsymbol{X}^*$ 充分融合，以提高特征表示质量，并将其用于跨视频伪标签指导模块。

参考文献

[1] CHENG F, BERTASIUS G. Tallformer: temporal action localization with long-memory transformer[C]//Proceedings of the 17th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2022: 503-521.

[2] TIRUPATTUR P, DUARTE K, RAWAT Y S, et al. Modeling multi-label action dependencies for temporal action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE Press, 2021: 1460-1470.

[3] ZHANG M, HU H Y, LI Z J. Temporal action localization with coarse-to-fine network[J]. IEEE Access, 2022, 10: 96378-96387.

[4] HE B, YANG X T, KANG L, et al. ASM-Loc: action-aware segment modeling for weakly-supervised temporal action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Press, 2022: 13915-13925.

[5] HU Y F, FU J, CHEN M Y, et al. Learning proposal-aware re-ranking for weakly-supervised temporal action localization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(1): 207-220.

[6] 郭文斌, 杨兴明, 蒋哲远, 等. 多时间尺度一致性的弱监督时序动作定位[J]. 计算机工程与应用, 2023, 59(10): 151-161.
GUO W B, YANG X M, JIANG Z Y, et al. Multi-temporal scales consensus for weakly supervised temporal action localization[J]. Computer Engineering and Applications, 2023, 59(10): 151-161. (in Chinese)

[7] 侯永宏, 李岳阳, 郭子慧. 基于对比学习的弱监督时序动作定位[J]. 天津大学学报, 2023, 56(1): 73-80.
HOU Y H, LI Y Y, GUO Z H. Weakly supervised temporal action localization based on contrastive learning[J]. Journal of Tianjin University, 2023, 56(1): 73-80. (in Chinese)

[8] ZHOU J X, WU Y. Temporal feature enhancement dilated convolution network for weakly-supervised temporal action localization[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Waikoloa, USA: IEEE Press, 2023: 6017-6026.

[9] HUANG L J, WANG L, LI H S. Weakly supervised temporal action localization via representative snippet knowledge propagation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Press, 2022: 3262-3271.

[10] QING Z W, SU H S, GAN W H, et al. Temporal context aggregation network for temporal action proposal refinement[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE Press, 2021: 485-494.

[11] SRIDHAR D, QUADER N, MURALIDHARAN S, et al. Class semantics-based attention for action detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE Press, 2022: 13719-13728.

[12] ZHENG Z, WANG P, LIU W, LI J, et al. Distance-IoU loss: faster and better learning for bounding box regression[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2020: 12993-13000.

[13] HUANG J, KONG M, CHEN L Y, et al. Temporal RPN learning for weakly-supervised temporal action localization[C]//Proceedings of the 15th Asian Conference on Machine Learning. New York, USA: PMLR, 2024: 470-485.

[14] LIU Y Y, ZHU H, REN H H, et al. Fusion detection network with discriminative enhancement for weakly-supervised temporal action localization[J]. Expert Systems with Applications, 2024, 238: 122000.

[15] PAUL S, ROY S, ROY-CHOWDHURY A K. W-TALC: weakly-supervised temporal activity localization and classification[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2018: 588-607.

[16] ZHANG C, CAO M, YANG D M, et al. CoLA: weakly-supervised temporal action localization with snippet contrastive learning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE Press, 2021: 16005-16014.

[17] GAO J Y, CHEN M Y, XU C S. Fine-grained temporal contrastive learning for weakly-supervised temporal action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Press, 2022: 19967-19977.

[18] LIU P, WANG C, QIN J, et al. Feature enhancement and foreground-background separation for weakly supervised temporal action localization[C]//Proceedings of the 5th ACM International Conference on Multimedia. New York, USA: ACM Press, 2024: 1-7

[19] QU S, CHEN G, LI Z, ZHANG L, LU F, KNOLL A. ACM-Net: action context modeling network for weakly-supervised temporal action localization[EB/OL]. [2021-04-07]. https://arxiv.org/abs/2104.02967.

[20] HONG F T, FENG J C, XU D, et al. Cross-modal consensus network for weakly supervised temporal action localization[C]//Proceedings of the 29th ACM International Conference on Multimedia. New York, USA: ACM Press, 2021: 1591-1599.

[21] JIANG H, TANG H Y, YAN M, et al. Revisiting unsupervised temporal action localization: the primacy of high-quality actionness and pseudolabels[C]//Proceedings of the 32nd ACM International Conference on Multimedia. New York, USA: ACM Press, 2024: 5643-5652.

[22] WANG C X, WANG J, XU W T. Double branch synergies with modal reinforcement for weakly supervised temporal action detection[J]. Journal of Visual Communication and Image Representation, 2024, 99: 104090.

[23] LI Z L, WANG Z L, LIU Q Y. Weakly supervised temporal action localization with actionness-guided false positive suppression[J]. Neural Networks, 2024, 175: 106307.

[24] YUN W, QI M, WANG C, MA H. Weakly-supervised temporal action localization by inferring salient snippet-feature[C]//Proceedings of the AAAI Conference on Artificial Intelligence. [S. l.]: AAAI Press, 2024: 6908-6916.

[25] YANG W F, ZHANG T Z, YU X Y, et al. Uncertainty guided collaborative training for weakly supervised temporal action detection[C]//Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, USA: IEEE Press, 2021: 53-63.
[26] LUO Z K, GUILLORY D, SHI B F, et al. Weakly-supervised action localization with expectation-maximization multi-instance learning [C] // Proceedings of the 16th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2020: 729-745.
[27] RADFORD A, KIM J W, HALLACY C, et al. Learning transferable visual models from natural language supervision[C]// Proceedings of the International Conference on Machine Learning (ICML). New York, USA: PMLR, 2021: 8748-8763.
[28] JU C, HAN T D, ZHENG K H, et al. Prompting visual-language models for efficient video understanding [C] // Proceedings of the 17th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2022: 105-124.
[29] LEI J, YU L C, BERG T L, et al. TVQA+: spatio-temporal grounding for video question answering[EB/OL]. [2019-04-225]. https://arxiv.org/abs/1904.11574.
[30] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, USA: IEEE Press, 2017: 4724-4733.
[31] FEICHTENHOFER C. X3D: expanding architectures for efficient video recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE Press, 2020: 200-210.
[32] ÇIÇEK Ö, ABDULKADIR A, LIENKAMP S S, et al. 3D U-Net: learning dense volumetric segmentation from sparse annotation [C] // Proceedings of the 19th Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). Berlin, Germany: Springer, 2016: 424-432.
[33] LI X, ZHONG Z S, WU J L, et al. Expectation-maximization attention networks for semantic segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea: IEEE Press, 2020: 9166-9175.
[34] ALDOUS D J. Lower bounds for covering times for reversible Markov chains and random walks on graphs[J]. Journal of Theoretical Probability, 1989, 2(1): 91-100.
[35] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation[C]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1532-1543.
[36] LIN Z J, ZHAO Z, ZHANG Z, et al. Weakly-supervised video moment retrieval via semantic completion network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. [S.l.]: AAAI Press, 2020: 11539-11546.
[37] IDREES H, ZAMIR A R, JIANG Y G, et al. The THUMOS challenge on action recognition for videos "in the wild"[J]. Computer Vision and Image Understanding, 2017, 155: 1-23.
[38] HEILBRON F C, ESCORCIA V, GHANEM B, et al. ActivityNet: a large-scale video benchmark for human activity understanding [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, USA: IEEE Press, 2015: 961-970.
[39] LIN T W, ZHAO X, SU H S, et al. BSN: boundary sensitive network for temporal action proposal generation[C]//Proceedings of the 15th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2018: 3-21.
[40] LONG F C, YAO T, QIU Z F, et al. Gaussian temporal awareness networks for action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA: IEEE Press, 2020: 344-353.
[41] ZHAO P S, XIE L X, JU C, et al. Bottom-up temporal action localization with mutual regularization [C] // Proceedings of the 16th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2020: 539-555.
[42] CHEN M Y, GAO J Y, YANG S C, et al. Dual-evidential learning for weakly-supervised temporal action localization[C]// Proceedings of the 17th European Conference on Computer Vision (ECCV). Berlin, Germany: Springer, 2022: 192-208.
[43] TANG X J, FAN J S, LUO C C, et al. DDG-Net: discriminability-driven graph network for weakly-supervised temporal action localization [C] //Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE Press, 2024: 6599-6609.
[44] REN H, YANG W F, ZHANG T Z, et al. Proposal-based multiple instance learning for weakly-supervised temporal action localization [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE Press, 2023: 2394-2404.
[45] ZHANG S C, ZHAO C H. Cross-video contextual knowledge exploration and exploitation for ambiguity reduction in weakly supervised temporal action localization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(6): 4568-4580.
[46] HUANG L J, WANG L, LI H S. Foreground-action consistency network for weakly supervised temporal action localization[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, Canada: IEEE Press, 2022: 7982-7991.
[47] LI J J, YANG T Y, JI W, et al. Exploring denoised cross-video contrast for weakly-supervised temporal action localization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans, USA: IEEE Press, 2022: 19882-19892.
[48] 曹雨欣. 弱监督时序动作增量定位方法研究[D]. 西安: 西安理工大学, 2024.
CAO Y X. Research on weakly supervised incremental local ization method for temporal actions [D]. Xi'an: Xi'an University of Technology, 2024. (in Chinese)
[49] ZHAO Y B, ZHANG H, GAO Z, et al. A snippets relation and hard-snippets mask network for weakly-supervised temporal action localization [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(8): 7202-7215.
[50] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C] //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1724-1734.
[51] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

文字编辑 金胡考
栏目编辑 赖玉玲