Computer Science and Application 计算机科学与应用, 2023, 13(4), 657-671
Published Online April 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.134065

Hans汉斯

两阶段的弱监督时序动作定位

骆文杰，江朝晖，单东风，史俊彪，熊思璇

合肥工业大学计算机与信息学院，安徽 合肥

收稿日期：2023年3月5日；录用日期：2023年4月3日；发布日期：2023年4月10日

摘　要

由于弱监督时序定位模型没有帧级的监督信号，模型识别动作实例在边界处容易出现两个问题：过多地关注动作最具识别的部分，忽略了动作的其他部分而导致了动作的欠定位；动作的边界处与背景极其相似，模型难以区分而导致了动作的过定位。为了进一步有效的分类动作片段，改善边界困难样本的欠定位和过定位问题，提出了一种两阶段的弱监督时序定位。该方法分为两个阶段，第一阶段中我们对输入的视频帧提取RGB和光流特征，设计一种困难样本挖掘策略，得到边界的困难样本集合和易动作样本集合。另外，我们设计了一种原型生成模块，得到了每个动作类别的原型中心，将第二阶段的动作分类任务转换成嵌入空间与原型中心的距离问题。在第二阶段中，输入第一阶段得到的困难样本集合，使用原型匹配模块得到特定的时间类激活图。另外光流特征因其表达动态的特性，应当给予重视。本文设计了一种困难样本集合与易动作样本集合进行相似度计算得到增强光流特征的方法，实现边界困难样本更加准确地动作预测。最后为了进一步优化模型预测的动作标签，采用伪标签策略，为模型提供有效的帧级监督信号。在THUMOS'14和ActivityNet v1.2数据集进行实验论证。实验结果表明，方法性能优于现有弱监督时序定位方法。

关键词

弱监督，动作定位，两阶段，原型学习，特征增强

Two-Stage Weakly Supervised Sequential Action Positioning

Wenjie Luo, Chaohui Jiang, Dongfeng Shan, Junbiao Shi, Sixuan Xiong

School of Computer Science & Information Engineering, Hefei University of Technology, Hefei Anhui

Received: Mar. 5th, 2023; accepted: Apr. 3rd, 2023; published: Apr. 10th, 2023

Abstract

Since the weakly supervised temporal localization model has no frame-level supervisory signal,

文章引用: 骆文杰, 江朝晖, 单东风, 史俊彪, 熊思璇. 两阶段的弱监督时序动作定位[J]. 计算机科学与应用, 2023, 13(4): 657-671. DOI: 10.12677/csa.2023.134065

the model recognizes action instances at the boundary and is prone to two problems: underlocalization of the action by focusing too much on the most recognized part of the action and ignoring the other parts of the action; overlocalization of the action by making the boundary of the action extremely similar to the background, which is difficult for the model to distinguish. In order to further classify action fragments effectively and improve the under- and over-localization problems of boundary-hard samples, a two-stage weakly supervised temporal localization is proposed. The method is divided into two stages. In the first stage, we extract RGB and optical flow features from the input video frames and design a difficult sample mining strategy to obtain the set of boundary difficult samples and the set of easy action samples. In addition, we design a prototype generation module to obtain the prototype center of each action category, and convert the action classification task in the second stage into a distance problem between the embedding space and the prototype center. In the second stage, the set of difficult samples obtained in the first stage is input and a specific temporal class activation map is obtained using the prototype matching module. In addition optical flow features should be given attention because of their property of expressing dynamics. In this paper, we design a method to obtain enhanced optical flow features by performing similarity calculation between the set of difficult samples and the set of easy action samples to achieve more accurate action prediction for boundary difficult samples. Finally, in order to further optimize the action labels predicted by the model, a pseudo-labeling strategy is used to provide an effective frame-level supervised signal for the model. Experimental demonstrations are performed on THUMOS'14 and ActivityNet 1.2 datasets. The experimental results show that the method performs better than existing weakly supervised temporal localization methods.

Keywords

Weakly Supervised, Action Localization, Two Stages, Prototype Learning, Feature Enhancement

1. 引言

时序动作定位的目的是在未裁剪的视频中找到动作开始和结束的位置。由于其在监视分析、视频汇总和检索等方面的广泛应用，学者们越来越重视时序动作定位的研究。传统上，强监督需要对每个视频中的每个动作实例进行起始和结束时刻的标注，这耗费了巨大的人力和算力，因此，只需要视频级标签的弱监督时序定位受到了越来越多的关注。弱监督时序定位的基本挑战在于视频动作分类和多个动作实例之间的映射关系，通过映射关系得到动作提议。现有的弱监督时序定位主要有两个分支，分别是基于多实例学习的和基于注意力机制的。基于多实例学习的机制[1] [2] [3] [4] [5]首先获得帧级动作分类分数，即类激活序列 CAS，然后使用 top-k 来选择包中的正例，构建视频级动作分类分数。另一种基于注意力的机制[6] [7] [8] [9] [10]直接从原始数据中预测帧级动作概率，将其作为注意力计算视频级的分类概率，从而得到模型的优化。然而这种方式存在着分类和检测之间的矛盾，即分类总是关注显著性高的片段，而检测应该不遗漏的发现整个动作实例，二者的矛盾导致了模型不好的表现。综上，由于缺乏帧级监督，两种方法都无法对前景和背景进行准确的分离，直接导致了模型不好的表现。

弱监督时序定位的前景是具有特定意图的连续动作模式，具有前景持续时间的变化范围大、前景和背景的边界处极其相似这两个特点。针对前景持续时间的变化范围大这一特点，我们发现模型预测的动作提议常出现欠定位的问题，即原本是动作的部分被误认为是背景，导致动作提议的割裂；针对前景和背景的

边界处极其相似这一特点，我们发现模型容易在动作边界处与背景混淆，得到动作的过定位，即原本是背景的部分被误认为是动作，导致提议的冗余。通过对样本的分析，我们发现动作的欠定位和过定位主要体现在两种困难样本中。如图 1 所示，一方面动作具有多个子动作，持续多个阶段，图 1(a)中显示跳远视频中，模型只关注起跳等有区别性的动作，而忽视了起跑、腾空相对固定时的动作部分，造成动作提议的欠定位；另一方面动作和背景的边界处十分相似，模型容易将背景误认为是动作，图 1(b)中显示在网球击球运动时，由于准备过程和击打网球时的动作特点高度相似，模型将准备部分和动作实例混淆，造成动作提议的过定位。根据图一的分析，我们分别将上述两种样本定义为假阴性样本和假阳性样本。

现有的方法对这两种困难样本提出了一系列的解决方法。一方面，针对假阴性样本，W-TALC [1]提出了共活动相似度损失来挖掘共同类别视频对的相似特征；CMCS [11]引入多样性损失来发现多个分支上的不同动作部分；AUMN [12]设计了一个记忆库来存储动作单元的外观和动作信息及其相应的分类器，之后在视频中寻找相同的动作单元来实现更完整的提议。另一方面，针对假阳性样本，TSCN [13]利用双流的后期融合可以消除不可靠的阳性提议来引导注意力，并且利用注意力进行迭代优化；ACSNet [14]引入带有显式的辅助上下文类别扩展标签，以促进动作 - 上下文分离的学习。然而这些方法都事先将视频分成 T 个片段，在片段内估计每帧的动作概率，没有利用到丰富的上下文关系来帮助困难样本的改善，其次输入的光流特征在运动信息上提供了更加重要的线索，不应与 RGB 特征有着同等的处理方式。当前的方法一般是采用早期融合的方法，将提取到的 RGB 特征和光流特征在输入网络之前就进行了拼接，大大减少了光流特征的作用。

通过分析以前方法的时间类激活图，我们发现假阴性样本和假阳性样本往往出现在动作的边界，并且是在梯度变化较大的位置，我们将其记为锚点位置。这些样本需要得到更加精细的分类。因此，本文提出了一种两阶段的弱监督时序定位方法。

首先我们设计了一种有效的边界困难样本挖掘策略，并且针对这些边界困难样本提出两种改善方式：

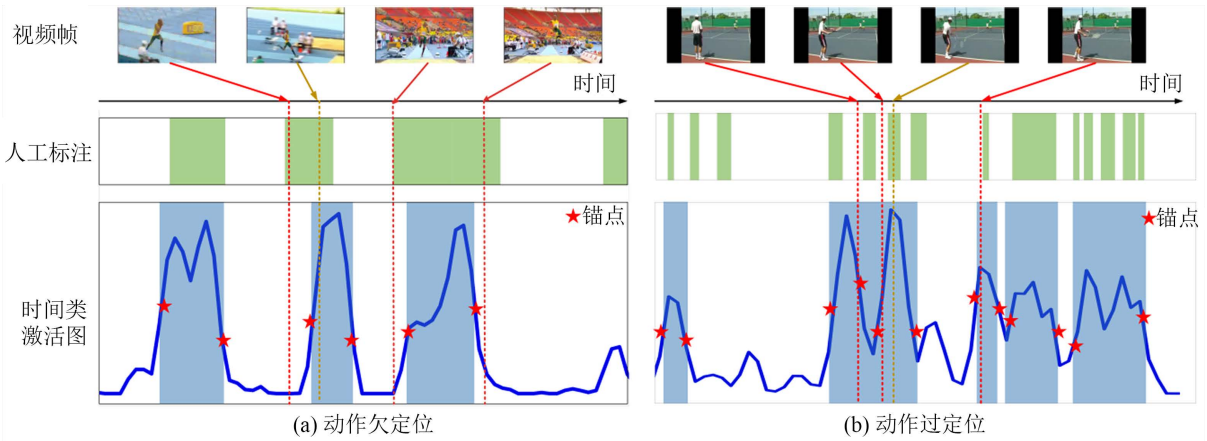


Figure 1. Two situations affecting the positioning effect of weak supervised timing sequence
图 1. 影响弱监督时序定位效果的两种情况

为了改善假阴性样本引起的动作提议的欠定位问题，我们发现如果同一动作的片段级特征尽可能相似，那么同一动作的片段在分类上就有相似的分类分数，提高动作提议的完整度。因此本文通过学习一种紧凑的特征表示，引入原型学习网络。为了改善由假阳性样本引起的动作提议的过定位问题，本文发现光流特征因其表达动态变化的特性，提出了一种光流特征单独处理，并且与动作响应程度大的帧做相似度计算，得到一种具有全局感知的光流增强特征。

本文的贡献如下：

1) 本文设计了一种两阶段的弱监督时序定位模型，对动作边界进行更加精细的分类。针对动作提议在动作边界的欠定位问题，我们提出了一种原型网络，将复杂的动作定位任务转换成原型中心与样本特征之间的距离；针对动作提议在动作边界的过定位问题，我们提出了额外关注光流特征，并且与响应程度大的动作帧做相似度计算，得到一种具有全局感知的光流增强特征。

2) 我们提出了一种困难样本挖掘算法，在边界附近定位潜在分类错误的样本，对这些样本进行更加精细的分类。

3) 在 THUMOS'14 [15]和 ActivityNet v1.2 [16]数据集上的大量实验证明了我们提出边界样本调优网络的有效性。

2. 方法

为了在弱监督环境下实现动作时序定位，本文设计了一个边界样本再分类网络。网络框架图如图 2 所示，它包含特征提取、困难样本挖掘模块、原型生成与使用模块、动作性模块、动作定位和损失函数。本文的核心思想在于对边界的困难样本进行挖掘、使用原型学习和动作性思想进行困难样本调优，最终达到边界困难样本的正确分类。

2.1. 特征提取

模型的输入是一组视频 $\{V_n\}_{n=1}^{N}$ 和它们对应的动作类别标签 $\{y_n\}_{n=1}^{N}$，其中 $y_n \in \mathbb{R}^C$，C 是动作类别的数目，如果第 n 个视频包含第 c_i 个动作类别，则 $y_{n,c}=1$，否则 $y_{n,c}=0$。对于每一个未剪辑的视频 V_n，我们将它分割成 n 个不重叠的片段 L_n，则 $V_n=\{S_{n,l}\}_{l=1}^{L_n}$。从时间长度上看，每一个片段 $\{S_{n,t}\}_{t=1}^{T}$ 的长度 T 是依赖于未裁剪视频的长度。然后通过预训练过的 I3D 模型，我们提取出每个片段的 RGB 特征 $X_n^R=\{x_t^R\}_{t=1}^{T}$ 和光流特征 $X_n^O=\{x_t^O\}_{t=1}^{T}$，其中 $x_t^R \in R^d$，$x_t^O \in R^d$，d 是每个片段的特征维度。最后，我们通过 f_{con} 方法将 RGB 特征和光流特征连接起来，构建更加紧凑的特征 $X_n \in \mathbb{R}^{T\times 2d}$，其中 f_{con} 是通过一个时序卷积串联一个 ReLU 激活函数组成。

为了预测片段级的动作分类得分，我们通过将特征 X_n^c 输入到两层时间 1D 卷积中，得到时间类激活图序列：

$$s^{base}=f_{cls}\left(X_n\right) \tag{1}$$

其中，$f_{cls}(\cdot)$ 表示分类器，具体包含两个 1D 卷积和 1 个 ReLU 激活函数，n 是视频样本的编号，时间类激活图 $s^{base} \in \mathbb{R}^{T\times(C+1)}$，$T$ 是视频中时间片段的长度，C 是动作类别编号，1 是我们引入背景类标签来解决样本和标签的不一致性。

为了判断每个片段是否为一个动作片段，并且为之后的困难样本挖掘做准备，我们对上述的时间类激活图在通道维度上进行聚合：

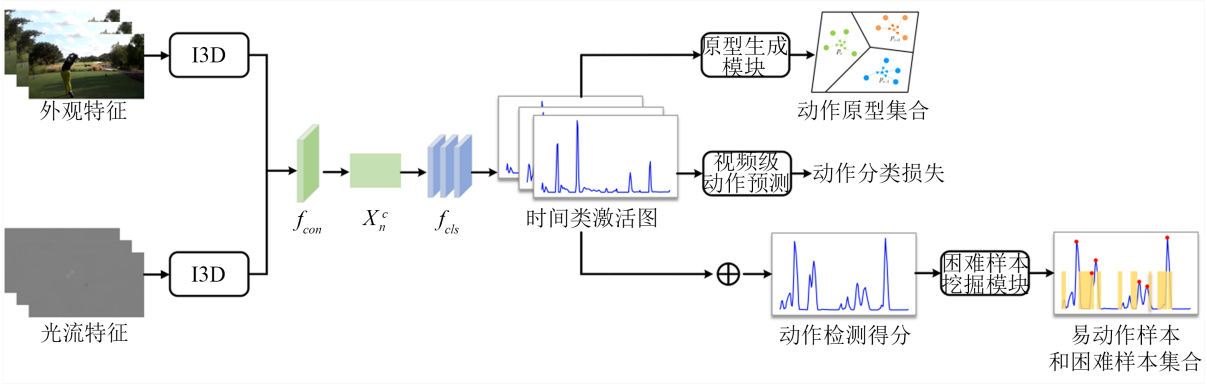
$$A_n^{ness}=sigmoid\left(f_{sum}\left(s^{base}\right)\right) \tag{2}$$

最终获得类无关的片段级动作检测得分 $A_n^{ness} \in \mathbb{R}^{1\times T}$。

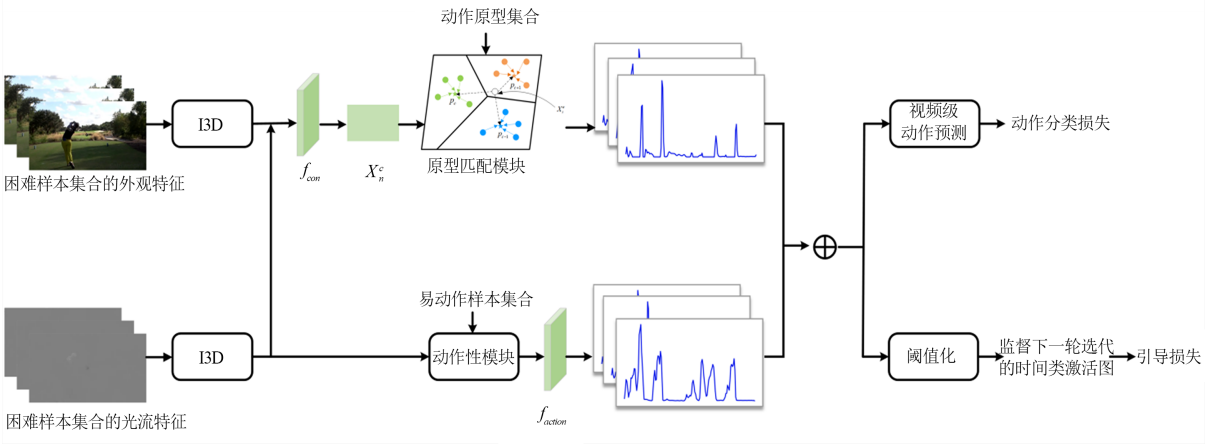
2.2. 锚点样本和困难样本挖掘

通过分析先前方法中的时间类激活图，我们发现影响动作定位的样本大多分布在动作的边界。直观

上看，对于大多数位于动作或者背景段内的样本，它们远离动作定位的边界，噪声干扰较少，在时间类激活图中有着较明显的响应程度；然而，对于动作定位边界相邻的样本，由于它们位于动作和背景之间的过渡区域，因此他们的可靠性较低，从而导致定位效果差。



(a)



(b)

Figure 2. (a) The first phase of this paper network; (b) the second phase of this paper network
图 2. (a) 本文网络第一阶段；(b) 本文网络第二阶段

本文设计了一种基于梯度的困难样本挖掘策略。通过分析数据集中动作实例的平均个数，首先设定超参数 θ^{anchor}，则一个视频被分成的片段个数为 $\frac{T}{\theta^{anchor}}$。

在每个片段内，通过最大值和最小值的方式找到动作响应度最大和最小的位置，对应动作的波峰 x_s^{peak} 和波谷 x_s^{valley}。之后我们在波峰和波谷之间搜索梯度变化最大的位置，将该位置作为我们挖掘困难样本的锚点：

$$x_s^{anchor}=\left\{x \mid A_n'^{ness}=\max\left(A_{n+1}^{ness}-A_n^{ness}\right), x\in\left[x_s^{valley}, x_s^{peak}\right]\right\} \tag{3}$$

其中，$A_n'^{ness}$ 表示第 n 帧动作响应的变化幅度，变化幅度越大越有可能是动作的开始或者结束，也就是模型最难分辨的位置。

基于上述公式，我们假设锚点位置是动作和背景的边界，即弱监督环境下潜在的困难样本中心点。因此，我们可以通过膨胀的方式来挖掘困难样本，公式如下：

$$\begin{cases} w^{sh} = \left(n_s^{peak} - n_s^{valley}\right) \cdot \theta^{sh} \\ v^{sh} = \left(n_{s+1}^{valley} - n_s^{peak}\right) \cdot \theta^{sh} \\ x_s^{hard} = \left\{x_n \mid n \in \left[n_s^{anchor} - w^{sh}, n_s^{anchor} + w^{sh}\right] \cup \left[n_s^{anchor} - v^{sh}, n_s^{anchor} + v^{sh}\right]\right\} \end{cases} \tag{4}$$

其中 x_s^{hard} 表示第 s 个片段中的困难样本集合，θ^{sh} 表示膨胀系数。如图 3 所示，红色区域是我们采集到的困难样本集合。第一阶段训练完成后，可以得到每个视频的困难样本集合 x^{hard} 和易动作样本集合 x^{peak} 。

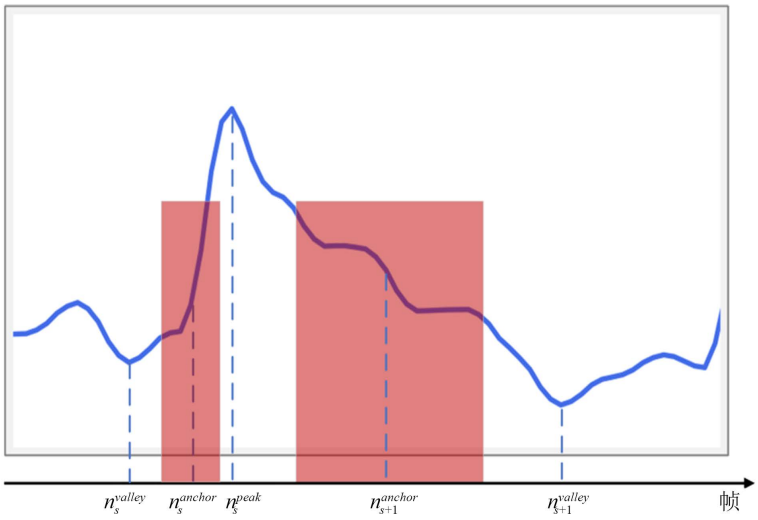


Figure 3. Diagram of difficult sample mining strategy
图 3. 困难样本挖掘策略的示意图

2.3. 原型生成与使用模块

原型学习用于依据动作类别的平均特征描述每个动作类，这在训练中标签较少时使用较多。在分类任务中，原型学习可以将复杂的特征分类任务，转换成嵌入空间中待分类样本和原型的距离匹配问题，所以原型学习的主要部分就是原型生成和原型匹配。

2.3.1. 原型生成模块

首先，随机初始化原型 $P^{(0)} = \left\{p_1^{(0)}, p_2^{(0)}, \cdots, p_c^{(0)}, p_{c+1}^{(0)}\right\}$，其中，$p_{c+1}^{(0)}$ 表示属于 $c+1$ 类别第 0 次迭代的原型特征。

其次，假设困难样本的数量共有 m 个，遍历所有的样本点 $i = 1, 2, \cdots, m$，根据原型与样本在空间上的欧氏距离，得到第 i 个样本属于第 j 个动作类别的概率，记为 $\mu_{ij} \in \mathbb{R}^{m \times (c+1)}$

$$\mu_{ij} = \frac{1}{\sum_{j=1}^{c+1} \left(\frac{\left\| x_i - p^{(t)} \right\|^2}{\left\| x_i - p_j^{(t)} \right\|^2} \right)^{\frac{1}{\gamma - 1}}} \tag{5}$$

其中 $p^{(t)}$ 表示第 i 个样本所属动作类别的原型中心，$p_j^{(t)}$ 表示第 j 个动作类别的原型中心，γ 表示迭代收敛程度。

然后，对每一个类别的原型 $p_1^{(t)}, p_2^{(t)}, \cdots, p_c^{(t)}, p_{c+1}^{(t)}$ 进行迭代更新，其中 t 为迭代次数。迭代公式如下所示：

$$p_j^{(t)} = \frac{\sum_{i=1}^{c+1} \mu_{ij}^{\gamma} x_i}{\sum_{i=1}^{c+1} \mu_{ij}^{\gamma}} \tag{6}$$

根据类紧凑性准则，收敛条件为所有样本点到该样本所在类中心的差异程度和最小：

$$p_j^{(t)} = \arg\min_j \sum_j^{c+1} \sum_{i=1}^{m} \mu_{ij}^{\gamma} \left\| x_i - p_j^{(t)} \right\|^2 \tag{7}$$

反复迭代，直到原型中心收敛。经过该算法流程，可以获得最终每个类别的动作原型集合 $P = \{p_1, p_2, \cdots, p_c, p_{c+1}\}$。

2.3.2. 原型匹配模块

根据原型网络，分类任务可以转化成用距离来衡量特征和原型之间的相似度，变成了原型匹配的过程。每个片段的相似度为：

$$s_{ij} = -\left\| x_i^{hard} - p_j \right\|_2^2 \tag{8}$$

其中 x_i^{hard} 表示第 i 个样本的嵌入特征，p_j 表示对应第 j 个类别的原型中心，$\|\cdot\|_2$ 表示 L2 正则化。之后将 s_{ij} 沿着类别维度输入到 softmax 中，得到每个样本对应类别的分类概率：

$$s_i^{proto} = p\left(x_i^{hard} \in p_j \mid x_i^{hard}\right) = \frac{\exp\left(s_j\right)}{\sum_{c=1}^{C+1} \exp\left(s_{i,c}\right)} \tag{9}$$

其中 $s_{i,c}$ 表示该样本与第 c 个原型中心的距离，s_j 表示该样本属于第 j 类动作类别的原型中心的距离。

与 STPN 中卷积层生成的时间类激活图一样，使用欧式距离生成的 $s^{proto} \in \mathbb{R}^{m\times(C+1)}$ 可以被视为用于动作定位的时间类激活图。

为了得到视频级的分类，我们将片段级的类激活图汇总。我们使用 topk 的方式来获得视频级分数：对于每一个类，我们取类激活图前 k 个最大值的项，并且计算它们的平均值，记为 a_c^{proto}，即视频属于动作类别 c 的视频级分数。

$$a_c^{proto} = \frac{1}{k} \sum_{i \in l} s_{i,c}^{proto} \tag{10}$$

其中 $s_{i,c}^{proto}$ 表示第 i 个样本对于第 c 类动作的得分，k 为超参数，控制视频中选择样本的数目，l 为得分最高的前 k 个片段集合，a_c^{proto} 表示视频对第 c 个动作类别聚合的分类得分。

对于视频级的分类得分，沿着类别维度应用 softmax 函数，使得视频级别的得分转化成属于每个类别的概率，公式如下：

$$\hat{y}^{proto} = \text{softmax}\left(a^{proto}\right) \tag{11}$$

其中，$\hat{y}^{proto} \in \mathbb{R}^{C+1}$ 表示该视频属于每个动作类别的可能性，总和为 1，$a^{proto} \in \mathbb{R}^{C+1}$，表示该视频属于每个动作类别的分类得分。

最后将视频级分类得分和视频级的动作类别标签做交叉熵损失函数，分类损失如下公式所示：

$$L_{cls}^{proto} = \sum_{c=1}^{C+1} -y_c \log\left(\hat{y}_c^{proto}\right) \tag{12}$$

其中 $\hat{y}_c^{proto}$ 为视频第 c 类的动作得分，y_c 表示视频第 c 类的动作标签，其中背景类别的标签为 0。

2.4. 动作性学习模块

传统的行为识别任务中，光流特征被广泛用于提供时间上的运动信息。但是光流特征的提取是使用

TV-L1 算法在连续的俩帧之间计算的，所以光流只能反映局部的运动信息。为了获得有效的运动信息，我们构建一个相似度学习模块，帮助光流学习全局的运动信息。

在未剪辑的视频中通常包含多个动作片段，因此寻找属于相同或者相似类别的动作实例将增强模型的判断能力。动作可能是在不同时间、不同场景、不同人物所发生，尽管这些时间距离较长，但是相同的行为特征让这些动作具有相似的语义信息。因此，我们选择第一阶段生成的易动作样本与困难样本建立联系，有助于为同一视频中相同或相似的动作实例建立全局的感知关系。假设困难样本的数目为 m，易动作样本的数目为 k，我们使用易动作样本帧与困难样本之间的余弦相似度 $\varepsilon^{sim} \in \mathbb{R}^{m\times k}$ 来判别语义相关的节点，公式如下：

$$\varepsilon_{ij}^{sim} = \frac{x_{i,flow}^{hard}\left(x_{j,flow}^{peak}\right)^{\mathrm{T}}}{\left\|x_{i,flow}^{hard}\right\|_2 \cdot \left\|x_{j,flow}^{peak}\right\|_2} \tag{13}$$

其中 ε_{ij}^{sim} 表示第 i 个困难样本的光流特征与第 j 个动作波峰样本的光流特征之间的相似度，$x_{i,flow}^{hard}$ 表示中的第 i 个困难样本的光流特征，$x_{j,flow}^{peak}$ 表示第 j 个动作波峰样本的光流特征，$\|\cdot\|_2$ 表示 L2 正则化。

之后我们将原本的锚点特征与相似度相乘，加权到原本的困难样本光流特征上，得到增强后的特征 $X_{action}^{hard} \in \mathbb{R}^{m\times d}$ 使得其拥有全局的感知能力，公式如下：

$$X_{action}^{hard} = X^{hard} + \alpha X^{peak} \cdot \varepsilon^{sim} \tag{14}$$

其中，α 为超参数，它起到平衡两种特征的作用。

在获得增强后的困难样本特征后，我们应用一个分类器来获得动作性得分 s_i^{action}：

$$s_i^{action} = f_{action}\left(x_{action.,i}^{hard}\right) \tag{15}$$

其中，$s_i^{action} \in \mathbb{R}^{m\times(C+1)}$ 表示第 i 个样本的动作性得分，分类器 $f_{action}(\cdot)$ 包含两个 1D 时序卷积和 ReLU 激活函数。

然后如公式(2)和(3)所示，使用 top-k 聚合得到视频级得分，利用 softmax 去获得动作性模块的每类动作的可能性，最后利用交叉熵损失去指导网络训练，如以下公式所示：

$$L_{cls}^{action} = \sum_{c=1}^{C+1} -y_c \log\left(\hat{y}_c^{action}\right) \tag{16}$$

其中 $\hat{y}_c^{action}$ 为视频第 c 类的动作得分，y_c 表示视频第 c 类的动作标签，其中背景类别的标签为 0。

2.5. 动作定位

模型最终的动作定位分成两个部分，包括第一阶段生成的非困难样本的动作定位和第二阶段生成的困难样本的动作定位。

针对第一阶段，如公式(2)和(3)所示，将第一阶段的时间类激活图 s_i^{base} 使用 top-k 聚合得到视频级得分，利用 softmax 去获得动作性模块的每类动作的可能性 $\hat{y}^{base}$。

针对第二阶段，在原型匹配模块和动作性模块之后，我们将两个分支使用平均聚合的方式合并成最终的时间类激活得分 $s^{total} \in \mathbb{R}^{m\times(c+1)}$，公式如下：

$$s_i^{total} = \frac{s_i^{proto} + s_i^{action}}{2} \tag{17}$$

其中 s_i^{proto} 为原型匹配模块的时间类激活得分，s_i^{action} 为动作性模块的时间类激活得分，s_i^{total} 为模型最终的时间类激活类得分。然后如公式(2)和(3)所示，使用 top-k 聚合得到视频级得分，利用 softmax 去获得

多时间尺度分支的每类动作的可能性。

得到第一阶段和第二阶段属于动作类别的概率后，我们使用分类阈值，过滤低于该阈值的动作类别。对于剩下的动作类别，我们使用一组阈值进行划分，生成动作提议。我们将第 i 个动作提议定义为 (b_i,e_i,c_i,q_i)，其中 b_i 为动作的开始时间，e_i 为动作的结束时间，c_i 为动作提议的类别，q_i 表示动作提议的置信度。得到所有的动作提议后，我们将动作提议集合输入到非极大抑制中，去除重复的提议，生成最终的视频定位结果。

2.6. 训练损失

该模型的损失共包含两个阶段，分别为第一阶段和第二阶段的损失。在第一阶段中，模型只需要视频级分类损失，公式如下：

$$L_{cls}^{base}=\sum_{c=1}^{C+1}-y_c\log\left(\hat{y}^{base}\right) \tag{12}$$

该模型的损失共包含两部分，分别为引导损失 L_{guide} 和视频级分类损失 L_{cls}。在第二阶段中，分支融合后的视频级分类损失公式如下：

$$L_{cls}^{total}=\sum_{c=1}^{C+1}-y_c\log\left(\hat{y}_c^{total}\right) \tag{18}$$

对于一致性损失，首先根据最终聚合的时间类激活得分，利用阈值 θ^{lab} 产生伪标签，高于阈值的认为是动作，置为 1；低于阈值的认为是背景，置为 0。具体如下：

$$g_{i,c}=\begin{cases}1, & s_{i,c}^{total}\geq\theta^{lab}\\ 0, & s_{i,c}^{total}<\theta^{lab}\end{cases} \tag{19}$$

其中 $g_{i,c}$ 表示第 i 个样本的伪标签，$s_{i,c}^{total}$ 表示第 i 个样本第 c 个动作类别的时间类激活得分。

当得到伪标签后，则强制每个模块得到的预测与伪标签相似，利用均方误差损失，如以下公式所示：

$$L_{guide}^{proto}=\sum_{i=1}^{m}\sum_{c=1}^{c+1}\left(s_{i,c}^{proto}-g_{i,c}\right)^2 \tag{20}$$

依据上述公式，同样可以得到动作性模块的引导损失 L_{guide}^{action}。

最终损失函数如公式所示：

$$\begin{cases}L=L_{cls}+\beta L_{\text{guide}}\\ L_{cls}=L_{cls}^{proto}+L_{cls}^{action}+L_{cls}^{total}\\ L_{guide}=L_{guide}^{proto}+L_{guide}^{action}\end{cases} \tag{21}$$

其中 β 为超参数，用来平衡损失。

3. 实验

3.1. 数据集与评价标准

在本节中，讨论数据集和实验设置的一些细节，之后与先前的方法进行了比较，对模型设计进行了消融研究，并在本节的最后展示了本文方法的可视化分析。

THUMOS’14 该数据集是弱监督时序定位广泛采用的数据集，由含有 20 个动作类别的未裁剪视频组成。训练集、验证集和测试集分别包含 13,320 个视频、1010 个视频和 1574 个视频。该数据集由于每个视频包含多个动作实例，每个视频包含多个动作类别，因此该数据集是具有一定挑战性的。按照[4]中常见的实验设置，我们使用验证集中的 200 个视频进行训练，使用测试集中 213 个视频进行测试。

ActivityNet 该数据集是弱监督时序定位的大规模数据集，包含两个版本：ActivityNet v1.2 和 ActivityNet v1.3。ActivityNet v1.3 由 200 个动作类别组成的 19,994 个未裁剪视频组成，训练、验证和测试的数据集比例为 2:1:1。

Table 1. Comparison with other methods on the THUMOS'14 dataset
表 1. 在 THUMOS'14 数据集上与其他方法的对比

监督方法	方法	mAP@IoU								
		0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
弱监督	W-TALC [1]	55.2	49.6	40.1	31.1	22.8	-	7.6	-	-
	BaS-Net [2]	58.2	52.3	44.6	36.0	27.0	18.6	10.4	3.9	0.5
	STPN [6]	52.0	44.7	35.5	25.8	16.9	9.9	4.3	1.2	0.1
	TSCN [13]	63.4	57.6	47.8	37.7	28.7	19.4	10.2	3.9	0.7
	UGCT (2021) [19]	69.2	62.9	55.5	46.5	35.9	23.8	11.4	-	-
	CoLA (2021) [20]	66.2	59.5	51,5	41.9	32.2	22.0	13.1	-	-
	UntrimmedNet [21]	44.4	37.7	28.2	21.1	13.7	-	-	-	-
	DGAM [22]	60.0	54.2	46.8	38.2	28.8	19.8	11.4	3.6	0.4
	本文方法	**70.2**	**66.3**	**56.0**	**46.8**	**36.4**	**24.3**	**12.0**	**4.2**	**0.9**
强监督	CDC [23]	-	-	40.1	29.4	23.3	13.1	7.9	-	-
	BSN [24]	-	-	53.5	45.0	36.9	28.4	20.0	-	-
	MGG [25]	-	-	53.9	46.8	37.4	29.5	21.3	-	-
	G-TAD [26]	-	-	54.5	47.6	40.2	30.8	23.4	-	-
	BC-GNN [27]	-	-	57.1	49.1	40.4	31.2	23.1	-	-
	P-GCN [29]	69.5	67.8	63.6	57.8	49.1	-	-	-	-

ActivityNet v1.2 是 ActivityNet v1.3 的一个子集，它涵盖了 100 个动作类别，训练集、验证集和测试集分别包含 4819 个视频、2383 个视频和 2480 个视频。每个视频平均还有 1.65 个动作实例。按照[4]中常见的实验设置，我们在训练集上训练模型，并在测试集上评估模型。

评价指标 本文采用与先前方法保持一致的评价指标，使用预测的动作提议和真实的动作区间在不同时间交并比(Intersection-over-Union, IoU)阈值下的平均精度阈值(mean Average Precision, mAP)作为评价指标，IoU 的阈值被设置为{0.1,0.2,⋯,0.9}。对于 ActivityNet v1.2，IoU 的阈值设置为{0.5,0.75,0.95}中选择，根据 ActivityNet 官方提供的针对时序动作定位任务代码评估模型的效果。

3.2. 实验细节

对于 THUMOUS 14 数据集，本文使用在 Kinetics [17]上预训练的 I3D 网络[17]作为特征提取器，并且为了公平比较并没有对视频特征进行微调。本文使用 RGB 和光流特征，其中光流特征是使用 TVL1 [18]算法提取的。之后将无重叠的连续 16 帧作为一个视频片段输入到 I3D 网络中。RGB 和光流特征经过 I3D 特征提取后，输出的特征维度都是 1024。

由于数据集中的未裁剪视频长度不同，本文将输入的视频片段调整成统一的长度。与先前方法[20]相同，在训练和测试时选择不同的采样策略，在训练时本文使用分层随机扰动的策略，在测试时使用均

匀采样的策略。本文所有实验使用的硬件配置为 Intel Core i7-5960X、CPU 3GHz 8cores RAM 8GB、图像显卡为 2 张 NVIDIA GeForce GTX 2080 Ti、Linux18.04 操作系统。软件框架使用 Pytorch 深度学习框架。训练时使用 Adam 优化算法学习率为 10^{-4}，权重衰减 5×10^{-4}。对于 THUMOS14 数据集，批处理大小设置为 16，一个迭代共有 4 轮，初始训练 100 个 Epoch，后续每轮迭代训练为 50 个 Epoch。对于 ActivityNet 1.2 数据集，批处理大小设置为 512，一个迭代共有 10 轮，初始训练 50 个 Epoch，后续每轮迭代训练为 20 个 Epoch。实验的超参数设置如下，α 和 β 的值分别为 0.5、0.25，θ^{lab} 为 0.5，θ^{sh} 为 0.5，θ^{anchor} 为 81。非极大值抑制(NMS)的阈值设置为 0.7。对于 THUMOS14 数据集 θ_{class} 设置为 0.25，T 为 750。对于 ActivityNet v1.2 数据集，θ_{class} 设置为 0.1，T 为 100。如果所有的动作类别得分均没有高于阈值的，选择得分最高的类别作为预测的动作类别。

3.3. 对比实验

表 1 展示了在此数据集上的对比实验结果。本文将对比方法主要分为两类：一种强监督，另外一种是弱监督。从表格中的数据可以看出，与弱监督的方法对比，本文方法的效果优于现有的弱监督时序定位方法，在 IoU = 0.5 时，mAP 达到了 36.9%。HAM-Net [8]提出了混合注意机制，包括时态软注意、半软注意和硬注意来关注动作的整个部分，而不是动作最具辨别的部分；UGCT [19]通过估计标签的不确定性来降低有噪声的伪标签误差；CoLA [20]通过对比学习挖掘困难片段，设计困难片段与容易片段间的对比损失来减少错误分类的片段。本文在 CoLA 的基础上，提出了一种新的困难样本挖掘策略，并且添加了原型学习模块和动作性模块，相比于其在 IoU = 0.5 时 mAP 提高了 4.7%。TSCN 使用双流后期融合的策略，可以避免单流中的假阳性样本干扰，但是他忽略了光流样本在判断动作性的特殊性，本文将光流特征进行相似性模块增强，得到了动作性特征，在 IoU = 0.5 时，mAP 提高了 8.2%。

与强监督的方法对比，MGG [25]结合基于锚点和基于边界的方法，准确生成动作提议；Long [26]等人引入高斯核来动态优化每个动作提议的时间尺度。本文方法在较低的 IOU 阈值情况下，与上述两个强监督方法差距不大，但是随着 IoU 阈值的增加，与强监督模型的差距增大。IoU 阈值越低，则表明预测正确的提议和真实的动作开始和结束的帧级注释差异越大，说明本文的模型可以知道动作大概发生的位置，但是更为准确的时间边界与强监督相比，还是有些差距。

Table 2. Comparison with other methods on the ActivityNet 1.2 dataset
表 2. 在 ActivityNet 1.2 数据集上与其他方法的对比

监督方式	Method	mAP@IoU			
		0.5	0.75	0.95	AVG
弱监督	BaSNet [2]	34.5	22.5	4.9	22.2
	STPN [6]	29.3	16.9	2.6	-
	HAM-Net [8]	41.0	24.8	5.3	25.1
	TSCN [13]	35.3	21.4	5.3	21.7
	CoLA [20]	42.7	25.7	5.7	23.6
	本文方法	**43.0**	**25.9**	**6.1**	**25.3**
强监督	BSN [24]	46.4	30.0	8.0	30.0
	BC-GNN [27]	50.6	34.8	9.4	34.2
	VSGN [28]	52.3	35.2	8.3	34.7

表2显示了在ActivityNet 1.2数据集中验证集和其他方法的对比结果，从表中数据可以看出，本文方法优于现有弱监督时序动作定位方法，其中AVG指的是IoU阈值从0.5到0.95，以0.05为步长的所有mAP的平均值。在IoU = 0.5时，与CoLA相比，提升从4.7%降到了0.3%，分析得出由于ActivityNet 1.2数据集的特点，即视频中的动作部分较多，所以两者的挖掘策略得到的片段几乎一致，所带来的提高有限。通过和其他方法的对比，验证了本文方法的有效性。

3.4. 消融实验

在本部分中，我们选择在 THUMOS14 数据集上进行消融实验分析，从而验证我们方法的有效性以及原型学习模块、动作性模块的效果。

首先，动作波峰样本和困难样本的数量对动作定位的结果是有影响的。表 3 给出了不同大小动作波峰样本数量在 IoU = 0.5 时 mAP 的值，表 4 给出了困难样本不同膨胀和内聚系数在 IoU = 0.5 时 mAP 的值。困难样本是根据动作波峰样本的增大而增大的，如果数量过少，容易导致过拟合的现象发生；如果数量过多，那么容易被分辨的样本被误以为是困难样本，降低第二阶段模型的鲁棒性，因此选择合适的动作波峰样本和困难样本对动作定位任务是有效的。

其次，我们通过对各个模块的组合来评估它们的贡献，如表 5 所示。基础模块为不使用第二阶段的原型学习和动作性模块，仅一次性的对全部样本做分类的模型。可以发现在添加原型学习模块后，IoU = 0.5 时 mAP 提高了，这说明使用原型学习模块可以得到更为准确的定位，改善了动作欠定位的问题，边界样本得到了精准分类。当仅使用动作性模块后，可以发现 IoU = 0.5 时，mAP 提高了，这说明给予光流特征更多的关注能有效减少误检的发生率。同时使用原型学习模块和动作性模块后 IoU = 0.5 时，mAP 提高了，效果更好。说明原型学习模块可以改善动作欠定位的问题，动作性模块可以改善背景混淆成动作的问题。

Table 3. mAP value when IoU = 0.5 is the number of action crest samples of different sizes
表 3. 不同大小动作波峰样本数量在 IoU = 0.5 时 mAP 的值

动作实例个数超参数 θ^{anchor}	1	3	27	81
mAP@0.5(%)	30.2	35.0	35.9	**36.4**

Table 4. mAP values of difficult samples with different expansion and cohesion coefficients at IoU = 0.5
表 4. 困难样本不同膨胀和内聚系数在 IoU = 0.5 时 mAP 的值

困难样本膨胀和内聚超参数 θ^{th}	0.1	0.25	0.5	0.75
mAP@0.5(%)	29.8	34.9	**36.4**	36.1

Table 5. The effect of the combination of modules on the model results
表 5. 各个模块的组合对模型结果的影响

基础模块	原型学习模块	动作性模块	mAP@IoU					
			0.1	0.2	0.3	0.4	0.5	AVG
✓			52.0	44.7	35.5	25.8	16.9	31.0
✓	✓		55.8	47.7	40.9	32.6	24.3	40.2
✓		✓	61.4	53.1	48.0	38.7	27.9	46.5
✓	✓	✓	**70.2**	**66.3**	**56.0**	**46.8**	**36.4**	**48.2**

3.5. 可视化分析

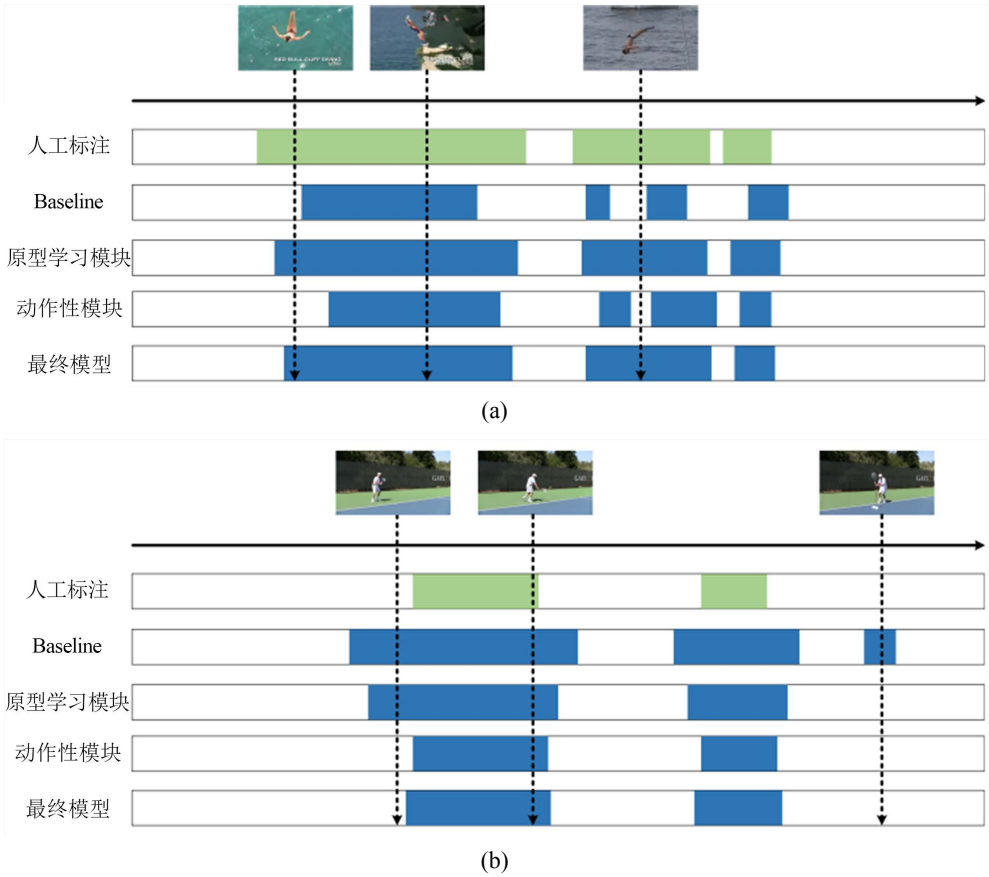


Figure 4. (a) Visualization results of diving actions; (b) visualization results of tennis strokes
图 4. (a) 跳水动作的可视化结果；(b) 网球击球动作的可视化结果

为了证明本文所提出方法的有效性，对于定位过程进行可视化分析，如图 4 所示。显示了三个具有代表性视频的动作定位结果。其中人工标注部分的绿色区域为动作片段，其余为背景片段；分类得分中的蓝色区域表示使用分类阈值后，得到的动作提议，即模型预测的结果。分类得分分为四个部分，分别为 baseline 分类得分、原型学习分类得分、动作性分类得分以及最终模型得分，其中最终模型得分为原型学习分类得分和动作性分类得分的平均值。各分类得分图的水平方向表示时间，垂直方向表示分类得分的大小。在图 4(a)的跳水样例中，本文方法解决了动作欠定位的问题。该样例有 3 个动作实例，其余都是背景片段。跳水动作持续时间很长，且有一些动作片段是在空中进行的，并不明显，很容易混淆成背景片段。对于 baseline，将在空中的动作被误认为是背景，导致了动作欠定位问题。在原型学习分类得分中，我们可以看到，模型很好的识别了这种空中动作，通过引入原型网络，学习到该动作类别的原型特征，将动作分类任务转换成嵌入空间的特征距离问题，帮助模型更加准确地识别了难以辨认的动作样本，证明了我们原型生成模块的有效性。在图 4(b)的网球击球样例中，本文方法解决了动作过定位的问题。该样例有 2 个动作实例，其余都是背景片段。由于网球击球前的准备片段与击球动作十分相似，这与上一个跳水动作实例不同，动作发生的速度变得更快了，很容易将准备片段误认为是动作。对于 baseline，将准备片段误认为是动作，导致了动作过定位问题。在动作性分类得分中，我们可以看到，模型准确地分辨出准备和击球动作的不同。动作性模块通过困难样本与易动作样本的相似性计算，并进行

特征增强，将准备样本的分类得分进一步降低，避免了基础模块中的错误判断。这也证实了我们将光流单独分析的有效性。同时，在图 4 可以看出，原型学习和动作性分支融合后的结果更加准确，这样得到的预测接近于人工标注，能够作为引导帮助网络训练，改善之后的分类错误。

4. 总结

本文提出了两阶段的弱监督时序定位方法。该模型使用困难样本挖掘算法，在边界附近定位潜在分类错误的样本，这些样本常使动作定位出现过定位和欠定位的问题，因此需要更加精细的分类。在第二阶段我们重新设计网络：针对动作提议的欠定位问题，我们提出了一种原型网络，将复杂的动作定位任务转换成原型中心与样本特征之间的距离；针对动作提议的过定位问题，我们提出了额外关注光流特征，并且与响应程度大的动作帧做相似度计算，得到一种具有全局感知的光流增强特征。两阶段模型通过训练后提高了动作定位的准确性，解决了动作提议的欠定位和过定位问题。对比实验证实了本文方法的有效性。后续工作可以考虑加入图卷积的思想，进一步挖掘全局的时间关系。

基金项目

安徽省重点研究与开发计划(202004d07020004)；安徽省自然科学基金项目(2108085MF203)；中央高校基本科研业务费专项资金(PA2021GDSK0072, JZ2021HGQA0219)。

参考文献

[1] Paul, S., Roy, S. and Roy-Chowdhury, A.K. (2018) W-TALC: Weakly-Supervised Temporal Activity Localization and Classification. *Computer Vision ECCV* 2018 15*th European Conference*, Munich, 8-14 September 2018, 588-607. https://doi.org/10.1007/978-3-030-01225-0_35

[2] Lee, P., Uh, Y. and Byun, H. (2020) Background Suppression Network for Weakly-Supervised Temporal Action Localization. 34*th AAAI Conference on Artificial Intelligence*, *AAAI* 2020, New York, 7-12 February 2020, 11320-11327.

[3] Moniruzzaman, M., Yin, Z., He, Z., *et al*. (2020) Action Completeness Modeling with Background Aware Networks for Weakly-Supervised Temporal Action Localization. *MM* '20: *The* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 2166-2174. https://doi.org/10.1145/3394171.3413687

[4] Luo, Z., Guillory, D., Shi, B., *et al*. (2020) Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning. *Computer Vision ECCV* 2020 16*th European Conference*, Glasgow, 23-28 August 2020, 729-745. https://doi.org/10.1007/978-3-030-58526-6_43

[5] Hong, F.T., Feng, J.C., Xu, D., *et al*. (2021) Cross-Modal Consensus Network for Weakly Supervised Temporal Action Localization. *Proceedings of the* 29*th ACM International Conference on Multimedia*, Chengdu, 20-24 October 2021, 1591-1599. https://doi.org/10.1145/3474085.3475298

[6] Nguyen, P., Han, B., Liu, T., *et al*. (2018) Weakly Supervised Action Localization by Sparse Temporal Pooling Network. 2018 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6752-6761. https://doi.org/10.1109/CVPR.2018.00706

[7] Narayan, S., Cholakkal, H., Khan, F.S., *et al*. (2019) 3C-Net: Category Count and Center Loss for Weakly-Supervised Action Localization. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 8678-8686. https://doi.org/10.1109/ICCV.2019.00877

[8] Islam, A., Long, C. and Radke, R. (2021) A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization. *The* 35*th AAAI Conference on Artificial Intelligence* (*AAAI*-21), 8-9 February 2021, 1637-1645. https://doi.org/10.1109/WACV45572.2020.9093620

[9] Huang, L., Wang, L. and Li, H. (2021) Foreground-Action Consistency Network for Weakly Supervised Temporal Action Localization. 2021 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 7982-7991. https://doi.org/10.1109/ICCV48922.2021.00790

[10] Nguyen, P.X., Ramanan, D. and Fowlkes, C.C. (2019) Weakly-Supervised Action Localization with Background Modeling. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October-2 November 2019, 5501-5510. https://doi.org/10.1109/ICCV.2019.00560

[11] Liu, D., Jiang, T. and Wang, Y. (2019) Completeness Modeling and Context Separation for Weakly Supervised Temporal Action Localization. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long

Beach, 15-20 June 2019, 1298-1307. https://doi.org/10.1109/CVPR.2019.00139

[12] Luo, W., Zhang, T., Yang, W., *et al*. (2021) Action Unit Memory Network for Weakly Supervised Temporal Action Localization. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 9964-9974. https://doi.org/10.1109/CVPR46437.2021.00984

[13] Zhai, Y., Wang, L., Tang, W., *et al*. (2020) Two-Stream Consensus Network for Weakly-Supervised Temporal Action Localization. *Computer Vision ECCV* 2020 16*th European Conference*, Glasgow, 23-28 August 2020, 37-54. https://doi.org/10.1007/978-3-030-58539-6_3

[14] Liu, Z., Wang, L., Zhang, Q., *et al*. (2021) ACSNet: Action-Context Separation Network for Weakly Supervised Temporal Action Localization. 35*th AAAI Conference on Artificial Intelligence* (*AAAI* 2021), Vancouver, 2-9 February 2021.

[15] Idrees, H., Zamir, A.R., Jiang, Y.-G., *et al*. (2017) The THUMOS Challenge on Action Recognition for Videos "in the Wild". *Computer Vision and Image Understanding*, **155**, 1-23. https://doi.org/10.1016/j.cviu.2016.10.018

[16] Heilbron, F.C., Escorcia, V., Ghanem, B., *et al*. (2015) ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 961-970. https://doi.org/10.1109/CVPR.2015.7298698

[17] Carreira, J. and Zisserman, A. (2017) Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 4724-4733. https://doi.org/10.1109/CVPR.2017.502

[18] Wedel, A., Pock, T., Zach, C., *et al*. (2009) An Improved Algorithm for TV-L^1 Optical Flow. *International Dagstuhl Seminar*, Wadern, 13-18 July 2008, 23-45. https://doi.org/10.1007/978-3-642-03061-1_2

[19] Yang, W., Zhang, T., Yu, X., *et al*. (2021) Uncertainty Guided Collaborative Training for Weakly Supervised Temporal Action Detection. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 53-63. https://doi.org/10.1109/CVPR46437.2021.00012

[20] Zhang, C., Cao, M., Yang, D., *et al*. (2021) CoLA: Weakly-Supervised Temporal Action Localization with Snippet Contrastive Learning. 2021 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 20-25 June 2021, 16005-16014. https://doi.org/10.1109/CVPR46437.2021.01575

[21] Wang, L., Xiong, Y., Lin, D., *et al*. (2017) UntrimmedNets for Weakly Supervised Action Recognition and Detection. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 6402-6411. https://doi.org/10.1109/CVPR.2017.678

[22] Shi, B., Dai, Q., Mu, Y., *et al*. (2020) Weakly-Supervised Action Localization by Generative Attention Modeling. 2020 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, 13-19 June 2020, 1006-1016. https://doi.org/10.1109/CVPR42600.2020.00109

[23] Shou, Z., Chan, J., Zareian, A., *et al*. (2017) CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1417-1426. https://doi.org/10.1109/CVPR.2017.155

[24] Lin, T., Zhao, X., Su, H., *et al*. (2018) BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 3-19. https://doi.org/10.1007/978-3-030-01225-0_1

[25] Yuan, L., Lin, M., Zhang, Y., *et al*. (2018) Multi-Granularity Generator for Temporal Action Proposal. *IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR* 2019, Long Beach, 16-20 June 2019, 3604-3613.

[26] Long, F., Yao, T., Qiu, Z., *et al*. (2019) Gaussian Temporal Awareness Networks for Action Localization. 2019 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 344-353. https://doi.org/10.1109/CVPR.2019.00043

[27] Yuan, H., Ni, D. and Wang, M. (2021) Spatio-Temporal Dynamic Inference Network for Group Activity Recognition. *Proceedings of the IEEE*/*CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 7476-7485. https://doi.org/10.1109/ICCV48922.2021.00738

[28] Zhao, C., Thabet, A.K. and Ghanem, B. (2021) Video Self-Stitching Graph Network for Temporal Action Localization. *Proceedings of the IEEE*/*CVF International Conference on Computer Vision*, Montreal, 10-17 October 2021, 13658-13667. https://doi.org/10.1109/ICCV48922.2021.01340

[29] Zeng, R., Huang, W., Gan, C., *et al*. (2019) Graph Convolutional Networks for Temporal Action Localization. 2019 *IEEE*/*CVF International Conference on Computer Vision* (*ICCV*), Montreal, 10-17 October 2021, 13638-13647. https://doi.org/10.1109/ICCV.2019.00719