Computer Science and Application 计算机科学与应用, 2024, 14(2), 183-199
Published Online February 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.142019

Hans汉斯

一致性对比采样网络的弱监督时序动作定位

陶应诚，黎　鑫，徐　浩，王　冠，景圣恩

合肥工业大学计算机与信息学院，安徽　合肥

收稿日期：2024年1月1日；录用日期：2024年2月1日；发布日期：2024年2月8日

摘　要

弱监督时序动作定位使用视频级标签，不需要高成本的动作实例标签，具有重要的研究价值。弱监督时序动作定位的难点在于，视频中的前景片段被淹没在背景片段中，难以得到精确的前景样本用于训练模型。关注于分析背景和前景片段在时间类激活序列上的差异，提出一致性对比采样网络。该网络使用多头注意力模块来增强行为特征。为了缓解前景样本被背景样本干扰的问题，该网络设计了易混淆样本的随机采样策略，用于学习前景采样的提议分布。为了促进前景分布的收敛，该网络联合考虑多阶段的前景采样规则，设计多阶段一致性采样模块。此外，针对前景和背景过渡区域的前景样本和背景样本较为相似，难以区分的问题，该网络设计对比采样模块，并联合考虑多阶段一致性采样，用于挖掘出困难前景样本，并使用对比学习优化困难前景样本的特征。在THUMOS 14和Activity v1.3数据集上进行实验验证。实验结果表明，提出的方法达到现有弱监督时序动作定位方法的性能。

关键词

时序动作定位，弱监督方法，一致性前景采样，对比采样

Consensus Contrastive Sampling Network for Weakly-Supervised Temporal Action Localization

Yingcheng Tao, Xin Li, Hao Xu, Guan Wang, Sheng'en Jing

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

Received: Jan. 1st, 2024; accepted: Feb. 1st, 2024; published: Feb. 8th, 2024

Abstract

Weakly supervised temporal action localization uses video-level labels, and does not require

文章引用: 陶应诚, 黎鑫, 徐浩, 王冠, 景圣恩. 一致性对比采样网络的弱监督时序动作定位[J]. 计算机科学与应用, 2024, 14(2): 183-199. DOI: 10.12677/csa.2024.142019

high-cost action instance labels. It has important research value. The difficulty of weakly supervised temporal action localization is that the foreground clips in the video are confused with the surrounding background clips, making it difficult to obtain accurate foreground samples for model training. This paper focuses on analyzing the difference between background and foreground clips on the temporal class activation sequence and proposes a consistent contrastive sampling network. The network uses a multi-headed attention module to enhance action features. To alleviate the problem that foreground samples are disturbed by background samples, the network designs a random strategy to sample confusable samples to learn the proposed distribution of foreground sampling. To facilitate the convergence of the foreground distribution, our network jointly considers multi-stage foreground sampling rules to design multi-stage consistent sampling modules. In addition, to address the problem that foreground and background samples in the foreground and background transition regions are highly similar and difficult to distinguish, our network designs the contrastive sampling module. Our network jointly considers multi-stage consistent sampling to select hard foreground samples and uses contrast learning to refine the features of hard foreground samples. Experiments on THUMOS 14 and Activity v1.3 datasets show that network achieves the performance of existing weakly supervised temporal action localization methods.

Keywords

Temporal Action Localization, Weakly-Supervised Method, Consensus Foreground Sampling, Contrastive Sampling

1. 引言

时序动作定位需要在未剪辑的视频中，确定行为发生的开始时间和结束时间，有利于发现行为发生的关键视频片段，被广泛应用于视频监控、视频摘要、行为检索等领域[1] [2] [3]。弱监督时序动作定位只需要视频级的标签，不需要高成本的动作实例标签。然而，在真实的未剪辑视频中，行为发生的前景片段常常淹没在大量的未发生行为的背景片段中，因此，弱监督时序动作定位受到背景噪声的严重干扰，是一个对输入数据敏感的病态问题。

弱监督时序动作定位的关键问题是如何实现背景噪声干扰下的建模问题。一个研究方向是包粒度的标签建模[4]，并使用多实例学习方法进行求解。其中，一个包对应一个视频，包中的一个实例对应一个行为片段，根据视频中的实例情况，将视频标记为正包或负包。一个正包至少包含一个正实例，而一个负包不包含正实例。多实例模型不仅可以预测视频标签，而且可以预测每个实例的标签。为了抑制背景实例，多实例方法采用池化方式来解决。在模型训练的反馈传播过程中，池化方法对所有实例的调整都是相同的，也就难以区分前景和背景实例，难以保证模型的收敛能力。另一个研究方向是片段粒度的伪标签建模[5] [6] [7]。片段粒度的网络模型通过学习类激活图得分来描述视频片段的行为，并利用该得分产生伪标签来计算损失函数。准确的类激活图得分依赖于可靠的前景样本分布。然而，上述方法没有充分分析大量背景样本干扰情况下的前景样本分布，因此，难以逼近真实的前景样本分布。

图 1 给出了罚球动作的前景持续时间以及前景片段采样策略。图 1(b)是基线模型的时间类激活序列。

图 1(c)是时序动作定位的前景片段人工标记。根据前景片段的人工标记可以看出，基线模型存在定位不完整和定位过度的问题。

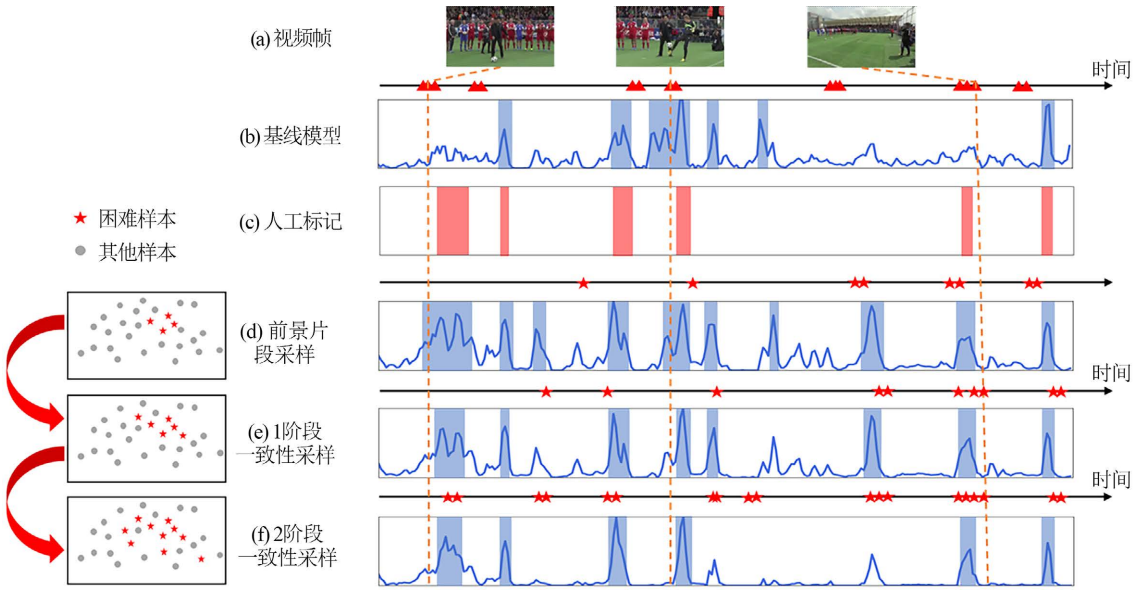

Figure 1. Foreground duration and foreground clip sampling strategy
图 1. 前景持续时间和前景片段采样策略

图 1(b)上方的时间轴上用红色三角形标出了对比采样中挖掘的困难样本，基线模型挖掘出了较多的困难样本，但多数困难样本最后都被分类为了背景，这是由于一次采样的局限性导致的。以图 1(b)上方时间轴三条虚线对应的三个困难样本为例。从左到右，第一个样本是足球教练演示罚球的动作并安排足球的位置，此时足球并未射向球门，可能是教练的动作导致了一个困难样本。第二个样本是守门员的颠球动作，而不是罚球射门的动作，足球的上下运动造成了干扰，这导致了一个困难样本。第三个样本是摄像机视角距离球员较远，足球的运动幅度在视频中较小，这导致了第三个困难样本。仅使用对比学习方法，是无法准确采样出所有前景样本的，因为有部分困难前景样本经过对比学习，会被判断为与背景样本极为相似，最终导致错误分类。为了避免这些困难前景样本被误分类为背景样本，需要进一步将这些易混淆的样本区分出来。

为了逼近真实的前景分布，本文根据弱监督前景样本的均值和方差设计易混淆样本的随机采样策略(图 1(d))。该策略假设样本的前景概率分布，在得分空间中是一种高斯分布的累积形式。该策略被设计为一种拒绝采样策略，即，均匀分布随机数通过累积高斯计算得到前景概率，如果视频帧的得分大于该前景概率，那么接受该样本，否则拒绝该样本。其中，均匀分布的随机数用于保证前景样本采样的多样性。为了进一步优化前景分布，避免一次采样的单一性，本文设计了一种多阶段的一致性随机采样策略(图 1(e)，图 1(f))，即在每个阶段都执行随机采样，来逐步过滤前景样本集合中的背景干扰样本。图 1(e)，图 1(f)是两个阶段的时间类激活序列，其上方的时间轴中标出了每一阶段中随机采样的样本，其左侧用红色五角星标出了采样得到的困难样本在二维空间中的分布情况。

由图 1(c)前景片段人工标记可以看出，不同的动作实例的持续时间是不同的，捕获大小相同时间感受野中的视频帧的依赖关系是不合理的，因此我们直接通过多头多尺度注意力来融合不同持续时间的动作实例的依赖关系，这有助于在源头上改善最终的定位效果。

一致性随机采样策略和易混淆样本的采样策略将背景干扰样本逐步剔除，为了优化筛选出的前景样

本特征，需要区分前景和背景过渡区域的相似样本，对比采样围绕前景片段的边界展开挖掘，优化出那些在边界上易被错误分类的前景样本。

本文的主要贡献如下：

1) 针对弱监督时序动作定位，视频包含大量与行为无关的样本的问题，本文设计了基于高斯提议分布的易混淆样本的随机采样策略。该策略通过调整样本的前景概率分布，实现多时间尺度下前景样本的有效筛选。

2) 针对现有单阶段时间类激活图采样，会限制前景样本采样准确性的收敛，本文设计了一种多阶段一致性前景采样策略。该策略联合考虑多个阶段一致性前景采样，进一步收敛采样准确性。

3) 针对在前景和背景过渡区域的前景样本和背景样本较为相似，难以区分的问题，本文方法联合考虑多阶段一致性采样和对比采样策略，用于挖掘出困难前景样本，并使用对比学习优化样本分布，提高前景样本和背景样本的可分离性。

2. 相关工作

2.1. 全监督时序动作定位

对未剪辑视频进行时序动作定位，全监督时序动作定位方法，主要采取两步策略，首先产生行为提议，然后对其进行行为分类。生成提议的方法，主要可以分为基于滑窗的方法和基于边界的方法。基于滑窗的方法是指滑动窗口或预定义锚点去产生固定尺寸的提议。行为上下文的图卷积网络[8]学习不同时间点之间的边界约束。一些方法关注于使用 Transformer 提取行为特征用于时序行为定位[9] [10]。全监督方法的主要缺点是人工标记成本高，难以应用到没有帧级标记的行为类别。

2.2. 弱监督时序动作定位

弱监督时序动作定位在训练时，仅仅需要视频级的行为标签，避免了全监督方法需要大量前景人工标记的问题。时间关系网络[11]使用 Transformer 来学习实例的时间关系特征。但是，这些工作仅在包粒度上进行分类和选择，难以区分前景和背景实例。

与注意力的软采样策略不同，删除样本是硬采样策略。深度片段选择网络[12]设计了多分支的样本删除策略。迭代优化定位方法[13]设计了多阶段的样本选择策略。时间域多分辨率特征方法[14]在多阶段框架内分析多尺度时间之间的特征约束关系。对比学习行为定位方法[15]关注于前景和背景的过渡的困难样本挖掘与特征优化。先验驱动模型[16]以动作特定场景先验、动作片段生成先验和可学习高斯先验的形式利用视频中潜在的时空规律，以补充现有的弱监督方法。区分性驱动的图网络[17]通过显式地模拟模糊片段和区别片段，防止模糊信息的传递，提高片段级表示的可区分性。但是，上述工作忽略了视频类激活图中，部分前景以低得分的方式存在的情况。

不同于现有工作，本文考虑提议分布来实现易混淆采样，并引入随机机制保留样本的多样性。该网络构建多阶段一致性采样模块，来保证前景分布建模的收敛，并使用对比学习来学习困难前景样本特征，用于提供前景样本和背景样本的可分离性。

3. 模型框架

图 2 描述了本文的一致性对比采样网络。该网络使用多尺度多头注意力模块来增强行为特征描述能力。基于提议分布的前景采样模块用于捕获前景样本分布。多阶段一致性采样模块通过联合考虑每个阶段的采样规则，来促使前景分布收敛。对比采样模块挖掘出困难的前景和背景样本，并使用对比损失进行易混淆样本的特征优化。

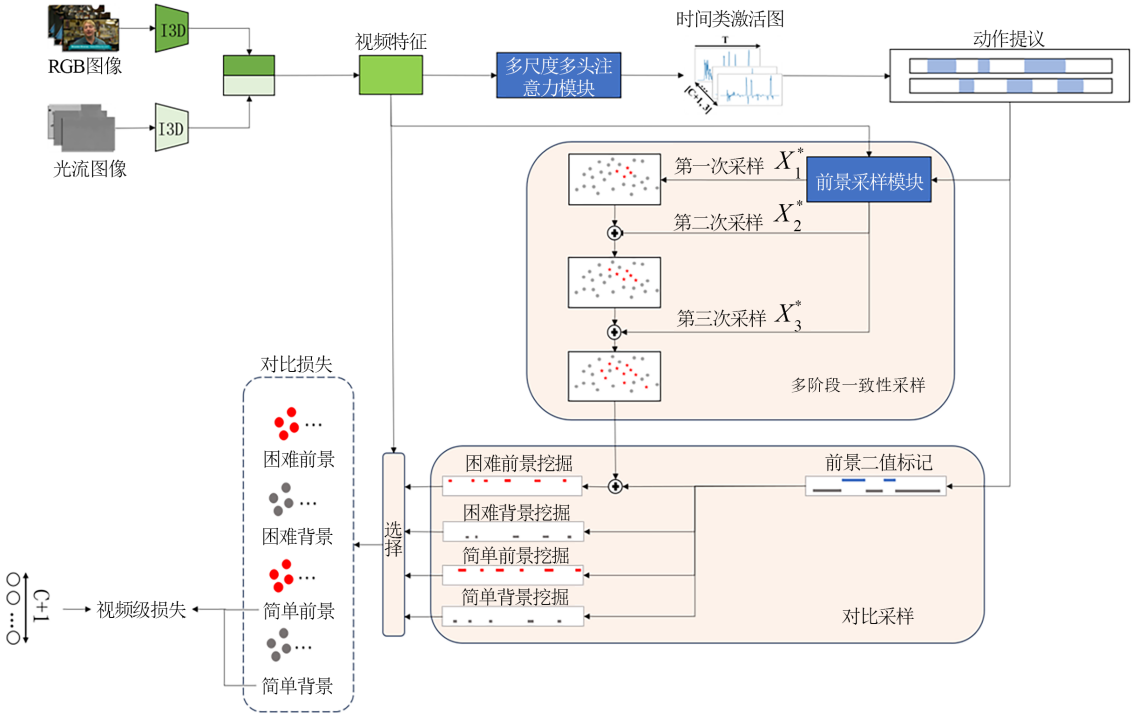

Figure 2. Consensus contrastive sampling network
图 2. 一致性对比采样网络

3.1. 多尺度多头注意力模块

多视频帧的上下文特征可用于区分行为类别。本文使用多头注意力来学习多视频帧中行为的时间依赖关系，以提取可靠的时间上下文特征。此外，行为具有不同的持续时间，单时间尺度特征难以同时捕获不同时间尺度的特征。因此，本文对三个时间尺度的特征分别使用多头注意力，以鲁棒提取不同持续时间的时间依赖特征。

每个视频被划分为 T 个视频片段。分别对该片段的 RGB 和光流图像，使用 I3D 网络进行特征提取。RGB 和光流特征串联后的特征维度为 $2D$。每个视频表示为一个 $T\times 2D$ 的特征矩阵 X。给定第 n 个片段的特征 x_n，使用两层时间 $1D$ 卷积和一层 *Sigmoid* 激活函数来学习第 j 个时间尺度的特征：$x'_{n,j}=\mathrm{Sigmoid}\left(conv\left(conv\left(x_n\right)\right)\right)$。其中，$j$ 是时间尺度的编号，其取值为 1，2，3。3 个分支卷积核的时间尺度分别为 1，3，5。

将一个视频的 N 个视频片段构建为样本集合。本文使用多头注意力来学习上下文特征。多头的编号为 $h=1,\cdots,H$。利用全连接层来学习每个注意力中的查询矩阵，关键值矩阵和值矩阵。第 h 个注意力的查询矩阵为 $Q_h=FC_h\left(x'_{1:N,j}\right)$，关键值矩阵为 $K_h=FC_h\left(x'_{1:N,j}\right)$，值矩阵为 $V_h=FC_h\left(x'_{1:N,j}\right)$。通过匹配查询矩阵和关键值矩阵，来学习片段之间的时间依赖关系。第 h 个注意力的特征为：

$$Att^h_{1:N,j}=FC_h\left(softmax\left(\frac{Q_h\cdot K_h^{\mathrm{T}}}{\sqrt{d}}\right)\cdot V_h\right)。$$

d 是 Q 的特征维度，即 $2D/h$，用于特征的尺度缩放，避免 *softmax* 后的特征进入梯度较小的范围。对多个头的注意力进行串联和线性变换，得到多头注意力特征为：

$$x''_{1:N,j}=Linear\left(concat\left(Att^1_{1:N,j},\cdots,Att^H_{1:N,j}\right)\right)。$$

最后，使用 Add 操作对时间尺度特征和上下文特征进行残差求和，并使用 Norm 操作对特征进行层归一化(*LN*: LayerNorm)，该归一化在单个 batch 中进行。多头注意力提取的时间类激活得分为：$r_{n,j}=LN\left(x'_{n,j}+x''_{n,j}\right)$。对每个视频片段，时间类激活得分是一个 $C+1$ 的特征向量，包括 C 个前景行为类别和 1 个背景行为类别。

3.2. 基于提议分布的前景采样模块

我们将训练视频拆分为样本集合 $X=\{x_i\}$，i 是训练视频片段的编号，其对应的时间类激活图得分集合为 $R=\{r_i\}$。并对每个前景类别进行后验概率 $h\left(y_c\mid x_i\right)$ 估计，其中 y_c 是第 c 个前景行为类别。由于真实的前景采样分布 $p\left(x_i\right)$ 复杂难以直接求解，容易造成样本是前景的后验概率 $h\left(y_c\mid x_i\right)$ 估计不准确。我们使用时间类激活图得分 r_i^c，来设计提议分布 $q\left(x_i\mid r_i^c,\theta\right)$，$\theta$ 为提议分布的参数。此时，前景后验概率的期望可以求解为：

$$\begin{aligned}E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]&=\int_{x_i}h\left(y_c\mid x_i\right)\cdot p\left(x_i\right)\mathrm{d}x_i\\&=\int_{x_i}h\left(y_c\mid x_i\right)\cdot\frac{p\left(x_i\right)}{q\left(x_i\mid r_i^c,\theta\right)}q\left(x_i\mid r_i^c,\theta\right)\mathrm{d}x_i\\&=E_{x_i-q\left(x_i\mid r_i^c,\theta\right)}\left[h\left(y_c\mid x_i\right)\cdot\frac{p\left(x_i\right)}{q\left(x_i\mid r_i^c,\theta\right)}\right]\\&=E_{x_i-q\left(x_i\mid r_i^c,\theta\right)}\left[h\left(y_c\mid x_i\right)\cdot\alpha\left(x_i\mid r_i^c,\theta\right)\right]\end{aligned}\tag{1}$$

其中，$\alpha\left(x_i\mid r_i^c,\theta\right)$ 是通过原始分布和提议分布的比值，引出了采样的权重。根据公式(1)可以看出当引入提议分布时，其原始分布的求解就可以转换到提议分布上进行，从而避免直接估计原始前景采样分布。即在实现提议分布时，等价于对样本权重的设计。本文设计了两种固定阈值的采样策略和两种高斯随机的采样策略。

低得分抑制的采样策略(*SS*: Strategy to Suppress samples with low score)，该策略对于得分低于阈值的不采样，高于阈值的保留为前景样本，有：

$$\alpha^{SS}\left(x_i\mid r_i^c,\theta\right)=\begin{cases}1 & r_i^c\geq u_c-\tau\\0 & r_i^c<u_c-\tau\end{cases}\tag{2}$$

其中，u_c 是弱监督行为类视频样本的得分均值，τ 是分数阈值参数。对该采样权重归一化，求得对应前景样本的概率分布为：

$$q^{SS}\left(x_i\mid r_i^c,\theta\right)=\begin{cases}1/\left(1+\tau-u_c\right) & r_i^c\geq u_c-\tau\\0 & r_i^c<u_c-\tau\end{cases}\tag{3}$$

易混淆样本的采样策略(*SC*: Strategy to sample Confusable ones)，该策略同时抑制低于低分阈值和高于高分阈值的样本，有：

$$\alpha^{SC}\left(x_i\mid r_i^c,\theta\right)=\begin{cases}0 & r_i^c>u_c+\tau\\1 & u_c-\tau\leq r_i^c\leq u_c+\tau\\0 & r_i^c<u_c-\tau\end{cases}\tag{4}$$

为了与低得分阈值策略对比，易混淆策略中使用相似的参数设计。对该采样权重归一化，求得对应的前景样本的概率分布为：

$$q^{SC}\left(x_i \mid r_i^c,\theta\right)=\begin{cases}0 & r_i^c > u_c+\tau \\ 1/(2\tau) & u_c-\tau \le r_i^c \le u_c+\tau \\ 0 & r_i^c < u_c-\tau\end{cases} \tag{5}$$

上述策略采用固定阈值采样，对前景样本的采样概率为 1，并没有考虑随机性，会限制模型的泛化能力。本文假设前景样本的分布在中心极限定理下逼近高斯分布，同时，由于高得分样本比低得分样本，有更大的概率是前景样本，因此，本文使用累积高斯分布来估计特定得分样本的可靠性，并引入拒绝采样来保持采样过程的多样性，从而将上述两种固定采样策略更新为随机采样策略。

低得分抑制的随机采样策略(*RSS*: Random Strategy to Suppress samples with low score)，该策略引入一个得分[0,1]区间内的均匀分布的随机数 z。如果得分大于该随机数则保留为前景采样，如果得分小于该随机数则拒绝采样。我们使用 $g\left(z \mid u_c,\sigma_c^2\right)$ 将 z 的均匀分布映射到高斯分布，有：

$$\alpha^{RSS}\left(x_i \mid r_i^c,\theta\right)=\begin{cases}I\left[r_i^c > g\left(z \mid u_c,\sigma_c^2\right)\right] & r_i^c \ge u_c-\tau \\ 0 & r_i^c < u_c-\tau\end{cases} \tag{6}$$

其中，$\left(u_c,\sigma_c^2\right)$ 是弱监督行为类视频样本的得分和方差，那么，样本随着随机变量 z 的采样概率为：

$$\begin{aligned}P\left(r_i^c\right) &> I\left[g\left(z \mid u_c,\sigma_c^2\right)\right]=\int_{-\infty}^{r_i^c} g\left(z \mid u_c,\sigma_c^2\right)\mathrm{d}z \\ &=\int_{-\infty}^{r_i^c}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left\{-\frac{\left(z-u_c\right)^2}{2\sigma_c^2}\right\}\mathrm{d}z\end{aligned} \tag{7}$$

对该采样权重归一化，求得对应的前景样本的概率分布为：

$$q^{RSS}\left(x_i \mid r_i^c,\theta\right)=\begin{cases}\dfrac{\int_{-\infty}^{r_i^c}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left\{-\frac{\left(z-u_c\right)^2}{2\sigma_c^2}\right\}\mathrm{d}z}{\int_{u_c-\tau}^{+\infty}\left(\int_{-\infty}^{r_i^c}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left\{-\frac{\left(z-u_c\right)^2}{2\sigma_c^2}\right\}\mathrm{d}z\right)\mathrm{d}r} & r_i^c \ge u_c-\tau \\ 0 & r_i^c < u_c-\tau\end{cases} \tag{8}$$

易混淆样本的随机采样策略(*RSC*: Random Strategy to sample Confusable ones**)**，该策略融合易混淆采样策略和基于高斯分布拒绝采样策略。对易混淆采样过程中，均匀分布的随机数 z，进行拒绝采样，有：

$$\alpha^{RSC}\left(x_i \mid r_i^c,\theta\right)=\begin{cases}0 & r_i^c > u_c+\tau \\ I\left[r_i^c > g\left(z \mid u_c,\sigma_c^2\right)\right] & u_c-\tau \le r_i^c \le u_c+\tau \\ 0 & r_i^c < u_c-\tau\end{cases} \tag{9}$$

同理，对该采样权重归一化，求得对应的前景样本的概率分布为：

$$q^{RSC}\left(x_i \mid r_i^c,\theta\right)=\begin{cases}0 & r_i^c > u_c+\tau \\ \dfrac{\int_{-\infty}^{r_i^c}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left\{-\frac{\left(z-u_c\right)^2}{2\sigma_c^2}\right\}\mathrm{d}z}{\int_{u_c-\tau}^{u_c+\tau}\left(\int_{-\infty}^{r_i^c}\frac{1}{\sqrt{2\pi}\sigma_c}\exp\left\{-\frac{\left(z-u_c\right)^2}{2\sigma_c^2}\right\}\mathrm{d}z\right)\mathrm{d}r} & u_c-\tau \le r_i^c \le u_c+\tau \\ 0 & r_i^c < u_c-\tau\end{cases} \tag{10}$$

我们对每个时间分支分别进行采样。此时视频样本 x_i，记作第 n 个视频片段第 j 个时间分支的视频样本 $x_{n,j}$。在第 j 个时间分支上，第 c 个行为类的高斯随机采样集合记作：$X^*_{n,j,c}=\left\{x_i\sim q^*\left(x_{n,j}\mid r^c_{n,j},\theta\right)\right\}$。其中，*={SS, SC, RSS, RSC}表示本文设计的四种采样策略之一。相对于单尺度随机采样，多尺度随机采样允许保留不同时间尺度的前景样本，具有更完备的前景特征分布捕获能力。

3.3. 多阶段一致性采样模块

在前景分布建模任务中，本文将后验概率的估计和采样策略的隐参数学习，看作是两个相对独立的过程，并使用期望最大化的思想，来求解采样策略中的参数。

第一步是求解期望，该过程核心在于求出前景的后验概率。由于分布的复杂性，期望无法直接计算，因此，我们使用深度模型学习的前景样本得分，来反映后验概率估计的可靠性。

第二步是最大化期望，该过程核心在于从后验概率中，求出采样隐参数 $\theta=\{\tau,\sigma_c\}$，即公式(3)中的分数阈值参数 τ，标准差参数 σ_c。由于弱监督样本中背景样本的存在，直接求解这两个隐参数并不能得到可靠的结果，因此，本文将这两个参数设置为超参数，来简化求解难度。

由于第一阶段时间类激活图得分采样的前景样本集合，其样本质量优于原始的弱监督样本，从而获得比原始弱监督样本更可靠的后验概率。为了进一步逼近真实分布，本文通过多次迭代的得分分析，来使得弱监督信号变强，逐渐逼近全监督信号，写作：

$$\lim_{k\to\infty}E_{x_i\sim q_k\left(x_i\mid r_i^c,\theta\right)}\left[h_k\left(y_c\mid x_i\right)\right]=E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]$$

或者写作：

$$\lim_{k\to\infty}E_{x_i\sim q_k\left(x_i\mid r_i^c,\theta\right)}\left[h_k\left(y_c\mid x_i\right)\right]-E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]=0$$

其中 s 表示迭代求解中的第 s 个阶段。对真实采样分布的求解过程，可以通过多阶段的提议分布逼近，具体的期望的迭代收敛过程可以写作：

$$\begin{aligned}
0&<E_{x_i\sim q_k\left(x_i\mid r^c_{k,i},\theta\right)}\left[h_k\left(y_c\mid x_i\right)\right]-E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]\\
&<E_{x_i\sim q_k\left(x_i\mid r^c_{k,i},\theta\right)}\left[h_{k-1}\left(y_c\mid x_i\right)\cdot\alpha_k\left(x_i\mid r^c_{k,i},\theta\right)\right]-E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]\\
&<E_{x_i\sim q_{k-1}\left(x_i\mid r^c_{k-1,i},\theta\right)}\left[h_{k-1}\left(y_c\mid x_i\right)\right]-E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]\\
&<E_{x_i\sim q_{k-1}\left(x_i\mid r^c_{k-1,i},\theta\right)}\left[h_{k-2}\left(y_c\mid x_i\right)\cdot\alpha_{k-1}\left(x_i\mid r^c_{k-1,i},\theta\right)\right]-E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]\\
&<\cdots\\
&<E_{x_i\sim q_1\left(x_i\mid r^c_{1,i},\theta\right)}\left[h_0\left(y_c\mid x_i\right)\right]-E_{x_i\sim p(x_i)}\left[h\left(y_c\mid x_i\right)\right]
\end{aligned}\tag{11}$$

在多阶段模型中，每个阶段学习到不同的特征。给定第 1 个阶段的特征 $r^c_{1,n,j}$，考虑所有样本 $n=1,\cdots,N$，所有时间尺度 $j=1,2,3$，所有行为类别 $c=1,\cdots,C$，则第 1 阶段，所有样本的采样集合为：$X^*_1=\left\{x_i\sim q^*\left(x_{n,j}\mid r^c_{1,n,j},\theta\right)\right\}$。其中，*={SS, SC, RSS, RSC}是采样策略。

在多阶段的采样过程中，本文引入一致性采样策略。其核心是第 s 阶段的采样时，需要满足第 1 阶段到第 s-1 阶段的所有条件，有利于逼近真实的前景分布。在第二阶段每个时间尺度上，采样的样本满足第 2 阶段得分的要求，或者满足第 1 阶段的抑制采样集合。一致性采样操作是上述两个集合的并集，此时，第 2 阶段用于损失计算的一致性采样集合为：$X^{*'}_2=\left\{x_i\sim q^*\left(x_{n,j}\mid r^c_{2,n,j},\theta\right)\bigcup x_i\in X^*_1\right\}$。其中，$n=1,\cdots,N$，$j=1,2,3$，$c=1,\cdots,C$。根据一致性采样操作，可以依次求出第 s 阶段的采样集合。

3.4. 对比采样模块

视频中的前景和背景的过渡区域存在高度相似的视频片段难以区分。因此，我们引入对比采样，来优化在过渡区域中的片段的特征。我们引入前景二值序列，来检测过渡区域。给定第 s 个阶段时间类激活图特征 $r_{s,n,j}^{c}$，我们对其所有行为类别求和，获得前景得分 $r_{s,n,j}^{ness}=Sigmoid\left(\sum_{c} r_{s,n,j}^{c}\right)$。将该得分与阈值 θ_b 比较，获得前景二值标记：

$$b_{s,n,j}=\begin{cases} fg & \text{if } r_{s,n,j}^{ness}\geq\theta_b \\ bg & \text{otherwise} \end{cases} \tag{12}$$

一个视频的所有片段，可以构建获得前景二值序列 $B_{s,j}=\left\{b_{s,n,j}\right\}$。根据该二值序列，我们将样本集合划分为困难前景样本，困难背景样本，简单前景样本，简单背景样本。前景二值序列二值序列可以腐蚀操作 $erode(.)$检测前景过渡区域 $X_{s,j}^{HA}$,记作 $erode\left(B_{s,j};m\right)-erode\left(B_{s,j};M\right)$其中，$m$ 是小的操作掩膜，M 是大的操作掩膜。困难前景采样集合是一致性采样集合和前景过渡集合的并集：
$X_{s,j}^{*,HA}=\left\{x_i\in X_{s,j}^{HA}\cup x_i\in X_{s}^{*}\right\}$。

前景二值序列可以使用膨胀操作 $dilate(.)$检测背景过渡区域 $X_{s,j}^{HB}$，记作 $dilate\left(B_{s,j};M\right)-dilate\left(B_{s,j};m\right)$。将该集合作为困难背景采样集合：$X_{s,j}^{HB}=\left\{x_i\in X_{s,j}^{HB}\right\}$。

对视频序列的得分进行降序排序，获得降序排序集合 $X_{s,j}^{DESC}$。选择得分较高的前 k^{easy} 个样本，并且不在过渡区域的样本作为简单的前景样本集合：$X_{s,j}^{EA}=\left\{x_i\in X_{s,j}^{DESC}\left[:,k^{easy}\right],x_i\notin X_{s,j}^{HA},x_i\notin X_{s,j}^{HB}\right\}$。

对视频序列的得分进行升序排序，获得降序排序集合 $X_{s,j}^{ASC}$。选择得分较低的前 k^{easy} 个样本，并且不在过渡区域的样本作为简单的前景样本集合：$X_{s,j}^{EB}=\left\{x_i\in X_{s,j}^{ASC}\left[:,k^{easy}\right],x_i\notin X_{s,j}^{HA},x_i\notin X_{s,j}^{HB}\right\}$。

3.5. 损失函数

第 s 个阶段，第 n 个视频片段，第 c 个行为类别的前景预测得分为：

$$p_{s,n,c}^{fg}=\frac{1}{3}\frac{1}{k^{easy}}\sum_{j=1}^{3}\sum_{k\in X_{s,j}^{EA}} r_{s,k,j}^{c}\cdot r_{s,k,j}^{ness} \tag{13}$$

我们估计背景得分为 1 减去前景得分，并作为权重来预测背景得分。第 s 个阶段，第 n 个视频片段，第 c 个行为类别的背景预测得分为：

$$p_{s,n,c}^{bg}=\frac{1}{3}\frac{1}{k^{easy}}\sum_{j=1}^{3}\sum_{k\in X_{s,j}^{EB}} r_{s,k,j}^{c}\cdot\left(1-r_{s,k,j}^{ness}\right) \tag{14}$$

本文方法的每个阶段是独立训练的，每个阶段只使用本阶段的损失函数。损失函数包括前景行为预测损失，背景行为预测损失，前景背景行为预测损失，和对比损失。

$$L_s^{total}=\lambda^{fg}L_s^{fg}+\lambda^{bg}L_s^{bg}+\lambda^{abg}L_s^{abg}+\lambda^{con}L_s^{con} \tag{15}$$

其中，$\lambda^{fg},\lambda^{bg},\lambda^{abg},\lambda^{con}$ 是超参数，用于平衡各种损失的作用。

前景行为预测损失用于鼓励视频级前景行为类别预测正确。该损失认为前景视频中不存在背景视频帧，将前 C 个类别设置给定视频级的标记 y，第 $C+1$ 个类别记为 0，前景行为的视频级标记记作 $y^{fg}=[y;0]$。对于训练集中的 N 个视频片段，前景行为预测损失计算为前景标记和前景预测得分的交叉熵损失：

$$L_s^{fg}=-\frac{1}{N}\sum_{n}\sum_{c=1}^{C+1} y_c^{fg}\cdot p_{s,n,c}^{fg} \tag{16}$$

背景行为预测损失用于鼓励非裁剪视频中存在背景片段。将前 C 个类别设置为 0 向量，第 $C+1$ 个类别记为 1，前景行为的视频级标记记作 $y^{bg}=[0;1]$。前景行为预测损失计算为背景标记和背景预测得分的交叉熵损失：

$$L_s^{bg}=-\frac{1}{N}\sum_n\sum_{c=1}^{C+1}y_c^{bg}\cdot p_{s,n,c}^{bg} \tag{17}$$

前景背景预测损失用于鼓励前景片段同时存在背景信息。例如，跳水行为和跳水台的背景信息是相关的，因此，背景信息不仅仅和背景预测相关，还和前景行为预测相关。将前景背景预测损失记作前景标记和背景预测得分的交叉熵损失：

$$L_s^{abg}=-\frac{1}{N}\sum_n\sum_{c=1}^{C+1}y_c^{fg}\cdot p_{s,n,c}^{bg} \tag{18}$$

对比损失用于优化困难前景/背景样本。为了优化困难前景样本，我们从困难前景集合中选择查询样本，从简单前景集合中选择正例样本，从简单背景集合中选择负例样本。为了优化困难背景样本，我们从困难背景集合中选择查询样本，从简单背景集合中选择正例样本，从简单前景集合中选择负例样本。我们计算样本之间的距离来描述样本分布，并采用交叉熵损失形式计算查询样本的对比损失：

$$loss^{con}\left(r_i,r^+,r^-\right)=-\log\left[\frac{\exp\left(\left(r_i\right)^T\cdot r^+/\theta_{con}\right)}{\exp\left(\left(r_i\right)^T\cdot r^+/\theta_{con}\right)+\sum_{k=1}^{k^{easy}}\exp\left(\left(r_i\right)^T\cdot r^-/\theta_{con}\right)}\right] \tag{19}$$

第 s 阶段，第 j 个时间尺度的对比损失包括困难前景样本和困难背景样本损失两个部分，有：

$$loss_{s,j}^{con}\left(r_i,r^+,r^-\right)=\begin{cases}loss_{r^+\sim X_{s,j}^{EA},r^-\sim X_{s,j}^{EB}}^{con}\left(r_i,r^+,r^-\right) & \text{if } r_i\sim X_{s,j}^{*,HA}\\ loss_{r^+\sim X_{s,j}^{EB},r^-\sim X_{s,j}^{EA}}^{con}\left(r_i,r^+,r^-\right) & \text{if } r_i\sim X_{s,j}^{HB}\end{cases} \tag{20}$$

多尺度的对比损失是每个尺度损失的平均值：

$$L_s^{con}=-\frac{1}{3}\frac{1}{N}\sum_j\sum_n loss_{s,j}^{con} \tag{21}$$

当训练完成后，我们使用前 k 个片段的得分来预测视频行为类别。我们分析大于视频分类阈值的类别进行时序行为定位。对于该行为类别，我们将大于片段阈值的片段作为候选片段。连续的候选片段用于产生行为提议。考虑非极大抑制策略，去除重复交并比高的提议，保留重复片段中得分高的提议。

4. 实验

4.1. 数据集

实验采用 THUMOS 14 数据集[18]，该数据集包含超过 24 个小时的 20 类不同行为的视频。其中训练集包含 2765 个已剪辑的视频，验证集包含 200 个未剪辑的视频，测试集包含 213 个未剪辑的视频。对于本文的时序动作定位任务，选择验证集用于训练视频，测试集用作测试。

实验同时使用 ActivityNet v1.3 数据集[19]，其中训练集有 10,024 个视频，验证集有 4926 个视频，测试集有 5044 个视频，包含有 200 个行为类。使用训练集视频用于训练，验证集用于测试。

4.2. 评价标准

本文使用 ActivityNet 官方提供的针对时序动作定位任务的方法，评估本文模型的效果。具体来说，

使用在不同时间交并比(IoU: Intersection of Union)阈值下的平均精度均值(mAP: mean Average Precision)作为评价标准。

4.3. 实验细节

网络使用在 kinetics 数据集上预训练的 I3D 模型提取视频特征。本文没有对 I3D 特征进行微调。两个流的特征维度都是 1024。光流提取使用 TVL1 算法。在 RGB 流和光流上都是将无重叠的连续 16 帧作为一个片段输入到 I3D 网络。THUMOS 14 数据集视频采样 750 个片段。ActivityNet v1.3 数据集视频采样 150 个片段。多头注意力的头的数量为 16。根据对比学习[15]，过渡区域检测中，小掩膜参数为 3，大掩膜参数为 6。样本的采样数量与视频片段数量相关。THUMOS 14 数据集的对比采样中，每个视频过渡区域的困难样本的采样数量为 23 个，简单样本的采样数量为 93 个。ActivityNet v1.3 数据集的对比采样中，每个视频过渡区域的困难样本的采样数量为 4 个，简单样本的采样数量为 6 个。困难前景样本是过渡区域的采样集合和一致性采样集合的交集。损失函数的超参数设置为 $\lambda^{fg}=1$， $\lambda^{bg}=0.5$， $\lambda^{abg}=0.5$，$\lambda^{con}=5\mathrm{e}-3$。

在时序行为定位过程中，THUMOS 14 数据集的视频分类阈值为 0.25，即允许视频中存在多个大于阈值 0.25 的类别被识别出。ActivityNet v1.3 数据集的视频分类阈值为 0.1。如果没有大于阈值的类别，选择得分最高的类别。计算 mAP 时，THUMOS 14 数据集使用多个片段阈值[0:0.15:0.015]，ActivityNet v1.3 数据集使用多个片段阈值[0:0.25:0.025]。对预测出的行为时间段进行非极大抑制，非极大值抑制的阈值为 0.7。

本文实验所采用的 PC 机配置为 Intel Core i7-5960X、CPU 3 GHz × 8 cores、RAM 8 GB，图像显卡为 1 张 NVIDIA GeForce GTX 1080Ti，Linux16.04 操作系统。深度学习框架为 pytorch。训练时使用 Adam 优化算法，学习率为 10~4，权重衰减 5x10-4。对于 THUMOS 14 数据集，每个阶段训练 200 个 epoch，批处理大小为 16。对于 ActivityNet v1.3 数据集，每个阶段训练 200 个 epoch，训练的批处理大小为 512。

4.4. 消融实验

4.4.1. 提议分布采样与对比采样

Table 1. Results of different sampling strategies on the THUMOS 14 dataset
表 1. 不同采样策略在 THUMOS 14 数据集的结果

	mAP@IoU (0.5)		
Baseline	32.6		
Baseline + 多头注意力	33.9		
Baseline + 多头注意力 + 对比采样	35.6		
+低得分抑制的采样策略	36.6		
+易混淆样本的采样策略	37.2		
随机采样策略/高斯标准差	0.05	0.1	0.2
+低得分抑制的随机采样策略	36.1	36.6	36.3
+易混淆样本的随机采样策略	36.7	37.2	36.8

表 1 给出了第 1 阶段，单尺度时间卷积核为 1 情况下，不同采样策略的 mAP@IoU (0.5)，即在 IoU 为 0.5 时的 mAP 结果。根据表 1 可以看出，1) baseline+多头注意力是指使用多头注意力的行为特征。该

特征提供的时间上下文特征，能够较好的改善前景样本的描述能力。2) 当本文使用低得分抑制的采样策略时，能够去除部分与行为无关的视频帧，从而改善模型结果。3) 当进行易混淆样本的采样策略时，去除了对模型学习优化能力不足的高分样本，使得模型学习对易混淆样本的损失信号更为敏感。4) 低得分抑制的随机采样策略，通过引入随机性来保证低分样本的多样性。当高斯方差较小时，其对低分样本的采样概率较小，当高斯方差较大时，其对低分样本的采样概率较大。当标准差为 0.1 的执行效果最好。5) 实验结果最好的实验条件是易混淆样本的随机采样策略，标准差为 0.1。此外，在相同的标准差下，易混淆样本随机采样策略的结果，优于低得分抑制的随机采样策略，这说明去除高分样本能够让模型更关注判决边界的学习，从而改善前景分布建模能力。

4.4.2. 多时间尺度

表 2 给出了两种随机采样策略，高斯标准差为 0.1 情况下，第 1 阶段不同时间卷积核的实验结果。表 2 说明，1) 在低得分抑制的随机采样策略中，结果随着 Conv1，Conv3，Conv5 依次提高。这说明当使用短时间卷积核时，难以考虑这些片段的上下文，对于运动变化不明显的行为，容易与背景混淆。同时，说明数据集中存在一些长的持续时间的行为片段，当引入长的时间卷积核时，能够描述更大范围的时间上下文，能够描述出行为在这个时间尺度上的变化，有利于前景从背景类中区分出来。2) 当 Conv7 添加到网络结构中，结果略微下降，这是由于 Conv7 估计更长时间段内的预测标记相同，这容易对夹在背景帧中的持续时间较短的行为进行误判。3) 在易混淆样本的随机采样策略中，最好的多时间融合策略是 Conv1 + 3 + 5，同样 Conv7 也会引起持续时间短的行为片段的误判。

Table 2. Results of different temporal convolution kernels on the THUMOS 14 dataset
表 2. 不同时间卷积核在 THUMOS 14 数据集的结果

	mAP@IoU (0.5)	
卷积核尺寸	低得分抑制的随机采样策略	易混淆样本的随机采样策略
Conv 1	36.6	37.2
Conv 1 + 3	37.7	38.3
Conv 1 + 3 + 5	39.1	39.6
Conv 1 + 3 + 5 + 7	38.8	39.3

4.4.3. 多阶段一致性采样

Table 3. Results of different execution phases on the THUMOS 14 dataset
表 3. 不同执行阶段在 THUMOS 14 数据集的结果

	mAP@IoU (0.5)	
执行阶段	低得分抑制的随机采样策略	易混淆样本的随机采样策略
1	39.1	39.6
2	40.7	41.3
3	41.9	42.5
4	42.4	43.0

表 3 给出了两种随机采样策略，高斯标准差为 0.1 情况下，Conv 1 + 3 + 5 的三个时间尺度下，不同阶段的结果。表 3 中迭代次数为 1 表示执行一次高斯随机采样。只有当迭代次数大于 2 时，才在每个阶

段训练完成后，考虑前驱各阶段的条件进行一致性采样。表 3 中可以看出，两种随机采样策略中，迭代次数 1，2，3 的结果依次提高。这说明在早期迭代过程中，能较好的删除行为类视频中与行为无关的视频帧，保留的样本能更好地逼近真实分布。迭代次数 4 的结果趋于收敛，模型趋于稳定。

4.5. 对比实验

Table 4. Results of state-of-the-art methods on the THUMOS 14 dataset
表 4. 现有方法在 THUMOS 14 数据集的结果

Method	mAP@IoU						
	0.1	0.2	0.3	0.4	0.5	0.6	0.7
全监督方法							
P-GCN (2019) [8]	69.5	67.8	63.6	57.8	49.1	-	-
TadTR (2022) [9]	-	-	74.8	69.1	60.1	46.6	32.8
ActionFormer (2022) [10]	-	-	82.1	77.8	71.0	59.4	43.9
弱监督方法							
BaS-Net (2020) [20]	58.2	52.3	44.6	36.0	27.0	18.6	10.4
DSSN (2021) [12]	57.8	49.7	41.0	32.3	22.4	-	-
MRITD (2021) [14]	61.2	55.5	47.1	38.5	29.7	20.1	11.5
ECM (2022) [21]	62.6	55.1	46.5	38.2	29.1	19.5	10.9
A-TSCN (2022) [22]	65.3	59.0	52.1	42.5	33.6	23.4	12.7
MSA (2021) [23]	65.5	58.9	49.1	40.0	31.4	-	-
CoLA (2021) [15]	66.2	59.5	51.5	41.9	32.2	22.0	13.1
AUMN (2021) [24]	66.2	61.9	54.9	44.4	33.3	20.5	9.0
TS-PCA (2021) [25]	67.6	61.1	53.4	43.4	34.3	24.7	13.7
ACM-Net (2021) [26]	68.9	62.7	55.0	44.6	34.6	21.8	10.8
ASM-Loc (2022) [7]	71.2	65.5	57.1	46.8	36.6	25.2	13.4
RSKP (2022) [27]	71.3	65.3	55.8	47.5	38.2	25.4	12.5
P-MIL (2023) [28]	71.8	67.5	58.9	49.0	40.0	27.1	15.1
CASE (2023) [29]	72.3		59.2		37.7		13.7
DDG-Net (2023) [17]	72.5	67.7	58.2	49.0	41.4	27.6	14.8
PivoTAL (2023) [16]	74.1	69.6	61.7	52.1	42.8	30.6	16.7
RSC	74.5	69.9	62.1	52.5	43.0	30.8	17.2

表 4 给出了 THUMOS 14 数据集上现有主流全监督方法的结果。Transformer 时序行为检测方法(TadTR) [9]考虑行为片段之间的位置来分析时间注意力。行为 Transformer (ActionFormer)方法[10]关注于多尺度 Transformer 来学习为特征。全监督方法的主要缺点是人工标记成本高，难以应用到没有帧级标记的行为类别。

表 4 给出了 THUMOS 14 数据集上的弱监督方法结果。CoLA 方法[15]关注于对比学习来优化困难行为特征。本文方法考虑多头注意力的行为特征。此外，本文方法考虑易混淆样本的随机采样策略(**RSC**)，并考虑一致性采样来增强行为样本分布建模，从而胜出 CoLA 方法。ASM-Loc 方法[7]考虑行为之间的时间关系。本文方法使用多头注意力考虑行为时间的时间关系，并考虑多阶段一致性采样和对比采样来优化行为特征，从而本文方法胜出 ASM-Loc 方法。DDG-Net 方法[17]考虑消除模糊信息的不利影响。本文考虑多阶段一致性采样策略，故本文方法胜出 DDG-Net 方法。

表 5 给出了 ActivityNet1.3 数据集上的弱监督方法的结果。背景标签分析是弱监督方法中的关键问题，融合标签分析的模型都取得了可靠的时序动作定位结果 CoLA [15]，UGCT [6]。本文方法采用一致性对比采样来分析前景样本，胜出现有的基于标签分析的弱监督方法。

Table 5. Results of state-of-the-art methods on the ActivityNet 1.3 dataset
表 5. 现有方法在 ActivityNet 1.3 数据集的结果

Method	mAP@IoU			
	0.5	0.75	0.95	AVG
BaS-Net (2020) [20]	34.5	22.5	4.9	22.2
DSSN (2021) [12]	34.3	21.0	6.1	21.5
ECM (2022) [21]	36.7	23.6	5.9	23.5
MRITD (2021) [14]	36.9	24.1	5.6	23.7
TS-PCA (2021) [25]	37.4	23.5	5.9	23.7
ACM-Net (2021) [26]	37.6	24.7	6.5	24.4
A-TSCN (2022) [22]	37.9	23.1	5.6	23.6
AUMN (2021) [24]	38.3	23.5	5.2	23.5
RSKP (2022) [27]	40.6	24.6	5.9	25.0
ASM-Loc (2022) [7]	41.0	24.9	6.2	25.1
P-MIL (2023) [28]	41.8	25.4	5.2	25.5
CASE (2023) [29]	43.2	26.2	6.7	26.8
PivoTAL (2023) [16]	45.1	28.2	5.0	28.1
RSC	45.8	28.9	6.5	29.2

4.6. 可视化分析

图 3 给出了两种随机采样策略的多阶段行为提议结果。在排球扣球中，该视频中只有一个行为示例。该行为中由于球超出拍摄高度，在低得分抑制随机采样策略的第 2 阶段出现了一个低得分片段。而实际上该片段中后方扣球人员从场景外跳起，并进入场景进行扣球。该片段的动作特点是垫球人员观察高空中的排球。本文低得分抑制采样策略，通过删除与行为无关的低得分样本，来发现该样本在内容上与扣球行为的相关性，从而在第 3 阶段找出了一个更长时间的行为。在易混淆样本随机采样策略中，同时删除低得分和高得分样本，来优化内容相关和内容无关样本的判决边界。在第 3 阶段中，易混淆样本随机采样策略的结果，在发现较长时间行为的同时，也去除了部分传球视频帧，这样检测结果能更关注扣球相关的行为。

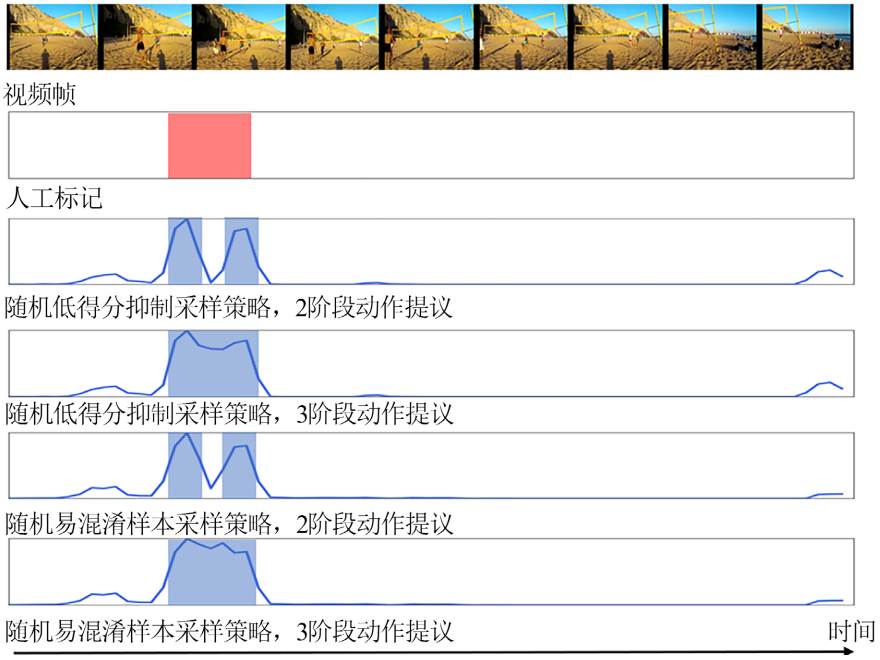

Figure 3. Multi-stage action proposal results with different strategies
图 3. 不同策略的多阶段行为提议结果

5. 总结与展望

本文提出一致性对比采样网络用于弱监督时序动作定位。该网络通过分析前景和背景在时间类激活图上的差异，设计基于提议分布的前景采样模块，来缓解弱监督时序动作定位中，前景样本被背景样本干扰严重的问题。本文设计了四种提议分布采样策略，分别是低得分抑制的采样策略，易混淆样本的采样策略，低得分抑制的随机采样策略，和易混淆样本的随机采样策略。为了促进前景分布的收敛，该网络联合考虑多阶段的前景采样规则，设计多阶段一致性采样模块。此外，针对前景和背景过渡区域的前景样本和背景样本较为相似，难以区分的问题，该网络设计对比采样模块，并联合考虑多阶段一致性采样，用于挖掘出困难前景样本，并使用对比学习优化困难前景样本的特征。在 THUMOS 14 和 Activity v1.3 数据集上的实验说明，易混淆样本的随机采样策略对前景样本提议估计是有效的。

参考文献

[1] Liu, Z., Wang, L., Tang, W., *et al.* (2021) Weakly Supervised Temporal Action Localization through Learning Explicit Subspaces for Action and Context. *Proceedings of the* 35*th AAAI Conference on Artificial Intelligence*, **35**, 2242-2250. https://doi.org/10.1609/aaai.v35i3.16323

[2] Islam, A., Long, C. and Radke, R. (2021) A Hybrid Attention Mechanism for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 35*th AAAI Conference on Artificial Intelligence*, **35**, 1637-1645. https://doi.org/10.1609/aaai.v35i2.16256

[3] 肖进胜, 申梦瑶, 江明俊, 雷俊峰, 包振宇. 融合包注意力机制的监控视频异常行为检测[J]. 自动化学报, 2022, 48(12): 2951-2959.

[4] Luo, Z., Guillory, D., Shi, B., *et al.* (2020) Weakly-Supervised Action Localization with Expectation-Maximization Multi-Instance Learning. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, Cham, 729-745. https://doi.org/10.1007/978-3-030-58526-6_43

[5] Zhai, Y., Wang, L., Tang, W., *et al.* (2020) Two-Stream Consensus Network for Weakly-Supervised Temporal Action Localization. In: Vedaldi, A., Bischof, H., Brox, T. and Frahm, J.M., Eds., *Computer Vision—ECCV* 2020, Springer, Cham, 37-54. https://doi.org/10.1007/978-3-030-58539-6_3

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

[7] He, B., Yang, X., Kang, L., *et al.* (2022) ASM-Loc: Action-aware Segment Modeling for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 2022 *IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 13915-13925. https://doi.org/10.1109/CVPR52688.2022.01355

[8] Zeng, R., Huang, W., Tan, M., *et al.* (2019) Graph Convolutional Networks for Temporal Action Localization. *Proceedings of the* 2019 *IEEE International Conference on Computer Vision*, Seoul, 27 October—2 November 2019, 7093-7102. https://doi.org/10.1109/ICCV.2019.00719

[9] Liu, X., Wang, Q., Hu, Y., *et al.* (2022) End-to-End Temporal Action Detection with Transformer. *IEEE Transactions on Image Processing*, **31**, 5427-5441. https://doi.org/10.1109/TIP.2022.3195321

[10] Zhang, C.L., Wu, J. and Li, Y. (2022) ActionFormer: Localizing Moments of Actions with Transformers. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M. and Hassner, T., Eds., *Computer Vision—ECCV* 2022, Springer, Cham, 492-510. https://doi.org/10.1007/978-3-031-19772-7_29

[11] Zhang, D., Huang, C., Liu, C. and Xu, Y. (2022) Weakly Supervised Video Anomaly Detection via Transformer-Enabled Temporal Relation Learning. *IEEE Signal Processing Letters*, **29**, 1197-1201. https://doi.org/10.1109/LSP.2022.3175092

[12] Ge, Y., Qin, X., Yang, D., *et al.* (2021) Deep Snippet Selective Network for Weakly Supervised Temporal Action Localization. *Pattern Recognition*, **110**, Article ID: 107686. https://doi.org/10.1016/j.patcog.2020.107686

[13] Pardo, A., Alwassel, H., Caba, F., *et al.* (2021) RefineLoc: Iterative Refinement for Weakly-Supervised Action Localization. *Proceedings of the* 2021 *IEEE Winter Conference on Applications of Computer Vision*, Waikoloa, 3-8 January 2021, 3318-3327. https://doi.org/10.1109/WACV48630.2021.00336

[14] Su, R., Xu, D., Zhou, L., *et al.* (2021) Improving Weakly Supervised Temporal Action Localization by Exploiting Multi-Resolution Information in Temporal Domain. *IEEE Transactions on Image Processing*, **30**, 6659-6672. https://doi.org/10.1109/TIP.2021.3089355

[15] Zhang, C., Cao, M., Yang, D., *et al.* (2021) Cola: Weakly-Supervised Temporal Action Localization with Snippet Contrastive Learning. *Proceedings of the* 2021 *IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 16010-16019. https://doi.org/10.1109/CVPR46437.2021.01575

[16] Rizve, M.N., Mittal, G., Yu Y, *et al.* (2023) PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 2023 *IEEE Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 22992-23002. https://doi.org/10.1109/CVPR52729.2023.02202

[17] Tang, X., Fan, J., Luo, C., *et al.* (2023) DDG-Net: Discriminability-Driven Graph Network for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 2023 *IEEE International Conference on Computer Vision*, Paris, 1-6 October 2023, 6599-6609. https://doi.org/10.1109/ICCV51070.2023.00609

[18] 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

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

[20] Lee, P., Uh, Y. and Byun, H. (2020) Background Suppression Network for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 34*th AAAI Conference on Artificial Intelligence*, **34**, 11320-11327. https://doi.org/10.1609/aaai.v34i07.6793

[21] Zhao, T., Han, J., Yang, L., *et al.* (2022) Equivalent Classification Mapping for Weakly Supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 3019-3031.

[22] Zhai, Y., Wang, L., Tang, W., *et al.* (2022) Adaptive Two-Stream Consensus Network for Weakly-Supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 4136-4151. https://doi.org/10.1109/TPAMI.2022.3189662

[23] Huang, L., Huang, Y., Ouyang, W. and Wang, L. (2021) Modeling Sub-Actions for Weakly Supervised Temporal Action Localization. *IEEE Transactions on Image Processing*, **30**, 5154-5167. https://doi.org/10.1109/TIP.2021.3078324

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

[25] Liu, Y., Chen, J., Chen, Z., *et al.* (2021) The Blessings of Unlabeled Background in Untrimmed Videos. *Proceedings of the* 2021 *IEEE Conference on Computer Vision and Pattern Recognition*, Nashville, 20-25 June 2021, 6172-6181. https://doi.org/10.1109/CVPR46437.2021.00611

[26] Qu, S., Chen, G., Li, Z., *et al*. (2021) ACM-Net: Action Context Modeling Network for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 35*th AAAI Conference on Artificial Intelligence*, **35**, 2233-2241. https://doi.org/10.1609/aaai.v35i3.16322

[27] Huang, L., Wang, L. and Li, H. (2022) Weakly Supervised Temporal Action Localization via Representative Snippet Knowledge Propagation. *Proceedings of the* 2022 *IEEE Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 3262-3271. https://doi.org/10.1109/CVPR52688.2022.00327

[28] Ren, H., Yang, W., Zhang, T. and Zhang, Y.D. (2023) Proposal-Based Multiple Instance Learning for Weakly-Supervised Temporal Action Localization. *Proceedings of the* 2023 *IEEE Conference on Computer Vision and Pattern Recognition*, Vancouver, 17-24 June 2023, 2394-2404. https://doi.org/10.1109/CVPR52729.2023.00237

[29] Liu, Q., Wang, Z., Rong, S., *et al*. (2023) Revisiting Foreground and Background Separation in Weakly-Supervised Temporal Action Localization: A Clustering-Based Approach. *Proceedings of the* 2023 *IEEE International Conference on Computer Vision*, Paris, 1-6 October 2023, 10433-10443. https://doi.org/10.1109/ICCV51070.2023.00957