

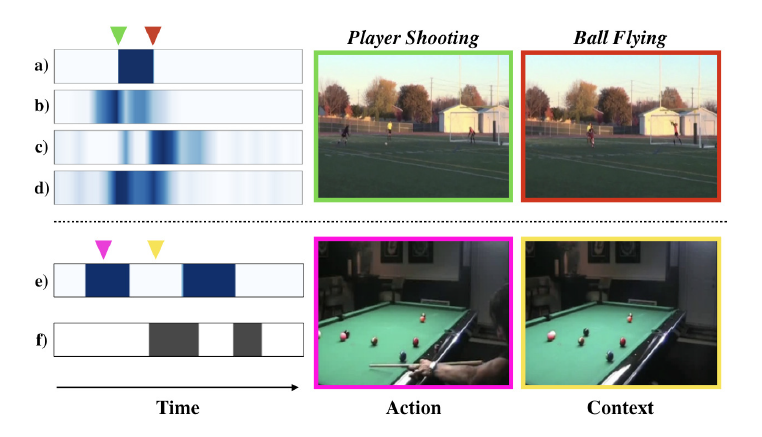


**弱监督时间动作定位的完整性建模和上下文分离**

## 摘要

*时间动作定位对于理解未修剪的视频至关重要。 在这项工作中，我们首先确定了对时间动作定位的弱监督带来的两个未充分探索的问题，即动作完整性建模和动作-上下文分离。 然后通过提出一种新颖的网络架构及其训练策略，明确地研究了这两个问题。 具体来说，为了模拟动作的完整性，我们提出了一个多分支神经网络，其中强制执行分支以发现独特的动作部分。 因此，可以通过融合来自不同分支的激活来定位完整的动作。 为了将动作实例与其周围的上下文分开，我们使用静止的视频剪辑不太可能是动作的先验来生成用于训练的硬负数据。 在 THUMOS’14 和 ActivityNet 数据集上进行的实验表明，我们的框架优于最先进的方法。 特别是 ActivityNet v1.2 上的平均 mAP 从 18.0% 显着提高到 22.4%。 我们的代码将很快发布。*

## 1 介绍



***Figure 1 说明这两个问题。 顶部：使用提议的多分支网络进行完整性建模。 训练分支以发现不同的动作部分，以便可以使用分支上的平均激活来定位完整的动作。 a) Soccer Penalty 动作实例的基本事实。 b) 我们训练模型中一个分支的类激活序列 (CAS)，定位玩家射击部分。 c) 另一个分支的 CAS，本地化 Ball Flying 部分。 d) 平均 CAS，完全定位实例。***

***底部：使用硬负视频生成(hard negative video Generation)进行上下文分离。 e) 台球动作实例的基本事实。 f) 获得的背景剪辑，区分同时出现的上下文。***

***右侧是取自相应颜色箭头的时间位置的四帧。 最好以彩色观看***

时间动作定位是一项重要的视觉任务，在视频监控 [42]、视频摘要 [28]、技能评估 [16] 等方面具有潜在应用。 目标是不仅要预测动作标签，还要从未修剪的视频中预测每个动作实例的开始和结束时间。 最近，完全监督的时间动作定位取得了显着进展[39、48、15、9、37、51、47、8、2、31]。 然而，精确注释动作实例的时间范围是劳动密集型和耗时的，这破坏了现实世界大规模场景中的完全监督方法。 因此，在训练期间只有视频级别的类别标签可用的弱监督设置更实用，并越来越受到社区的关注。 本文研究具有此类弱标签的时间动作定位。

大多数现有的弱监督方法 [45, 33, 38, 36, 52] 都属于多实例学习 (MIL) [54] 的框架。 在这个框架中，视频被视为一袋采样帧或片段，并输入视频级分类网络。 然后使用类激活序列 (CAS) [38] 对动作实例进行定位，这是每个动作的一维时间分类得分序列。

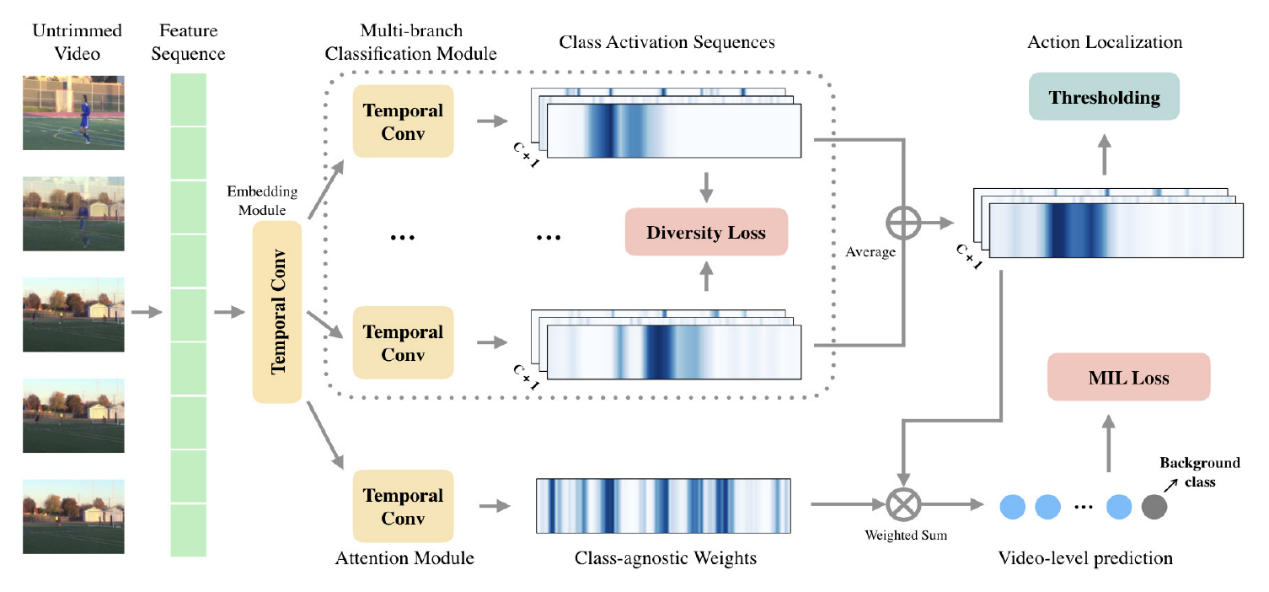
与其完全监督的对应物相比，弱监督的时间动作定位引入了两个新的挑战，称为**动作完整性建模和动作上下文分离**。 这两个问题之前没有得到很好的考虑，并且明显限制了性能。 第一个挑战是如何在没有完整注释的情况下完整地检测动作实例。 动作本质上是基本子动作的时间组合[20]，它们应该完全包含在预测中而没有遗漏。 在完全监督的环境中，一个动作是否完整是直接从时间边界的基本事实中学习的。 相比之下，当监督较弱时，由于定位任务现在被表述为视频级别的分类，因此缺乏细粒度的注释会使完整性建模复杂化。 识别动作的一个片段足以用于视频级分类，但不适用于片段级定位。 例如，Soccer Penalty 动作大致可以分为两个子动作，即 Player Shooting 和 Ball Flying。 仅在更具辨别力的 Player Shooting 部分上的激活足以对视频进行分类，但会使 Ball Flying 部分对定位产生假阴性。

动作-上下文分离的第二个挑战是如何将动作实例与带有弱标签的上下文区分开来。 同一类的动作实例通常被视觉上相似的剪辑包围，例如动作台球通常由屏幕中带有静态台球桌的评论剪辑包围。 这些片段与大多数视频中的真实动作一起出现，因此在本文中称为上下文。 上下文剪辑与普通背景剪辑在视频中的分布不同。 上下文剪辑在大多数情况下与真实动作同时出现，并且不涉及其他动作类别的视频，而背景剪辑是与类别无关的并且随机分布的。出于这个原因，上下文剪辑可以被视为硬底片。 视频级分类器学习具有相同标签的视频之间的相关性并发现它们的共同内容，不幸的是，这些内容不仅包括共同的动作（例如台球），还包括共同的上下文（例如静态台球桌）。 我们认为，在监督薄弱的情况下，动作-上下文分离本质上是困难的，除非使用关于动作的先验知识。

为了分别解决这两个问题，我们提出了多分支网络架构和硬负(hard negative)数据生成方案。 为了模拟动作完整性，从输入视频中提取的特征序列被馈送到具有多个并行分类分支的网络中。 设计了多样性损失以确保不同分支输出的类激活序列之间的差异，以便训练每个分支以定位动作的不同部分。 因此，如图 1 中的示例，可以通过聚合来自多个分支的激活来检索完整的操作。 然后随着时间的推移将类激活与时间注意力集中起来，产生视频级的类别分布。 我们使用基本事实计算其交叉熵，即标准 MIL 损失，该损失与多样性损失一起被最小化以学习网络参数。 至于动作-上下文分离，我们开发了一种简单而有效的策略，使用动作应该是运动的先验来挖掘硬底片。 我们在训练视频中搜索固定剪辑，如图 1 所示。然后使用静态剪辑生成伪视频并用新的背景类标记。 只要生成的伪视频中包含一些硬否定，这样的策略就可以帮助模型拒绝公共上下文。

在两个基准数据集 THUMOS'14 [21] 和 ActivityNet [6] 上，所提出的方法优于最先进的方法，证明了处理这两个问题的有效性。 总之，我们的贡献有三方面：1）提出了一个具有多样性损失的多分支网络来模拟动作完整性。 2) 设计了一种硬否定视频生成方案来分离公共上下文。 3）我们的方法在两个基准数据集上取得了优异的结果

### 2 相关工作



***Figure2 概述架构。 所提出的多分支网络由特征提取模块、特征嵌入模块、多分支分类模块和时间注意(temporal attention)模块组成。 在分类模块中，使用多样性损失训练多个分支以发现不同的动作部分。***

过去对修剪视频的动作识别进行了广泛的研究。 早期的方法主要基于手工制作的特征[26,43,34]。 近年来，已经提出了各种深度网络，例如双流网络[40、46]、LSTM [12]、3D ConvNets [41]、I3D [7]等[23、44、53]。 请参阅最近的调查 [1, 3, 22, 19] 进行详细回顾。

**完全监督的时间动作定位方法**主要基于提议加分类范式 [39, 37, 51, 15, 5, 9, 47, 8, 31]，其中首先生成时间提议然后分类。还研究了其他类别的方法，例如基于单次检测器 [4, 30] 或顺序决策过程 [48, 2] 的方法。 给定完整的注释，proposal-plus-classification 方法通常通过二元动作分类器在proposal阶段过滤掉公共上下文。 至于完整性建模，赵等人。 [51] 使用结构时间金字塔池，然后是显式二元分类器来确定实例是否完整。 侯等人。 [20]将一个动作的视频片段聚类成不同的子动作，然后将整个动作检测为子动作的有序序列。 袁等人。 [49] 将一个动作分为三个部分，即开始、中间和结束，以模拟其时间演变。 但它们都需要完整的注释。其他关于时空动作检测 [17] 和视频时间分割 [27] 的工作超出了我们的范围。

**弱监督时间动作定位算法**主要属于多实例学习（MIL）[54]。 王等人。 [45] 提出了一个名为 UntrimmedNet 的框架，由一个分类模块和一个选择模块组成，随后在 [33] 中引入了稀疏正则化。 保罗等人。 [36] 使用共活性相似性损失来强制同一类的局部实例之间的特征相似性。 AutoLoc [38] 不是在 CAS 上设置阈值，而是直接预测时间边界以检测动作。 关于第一个挑战，有两个先前的工作试图对动作完整性进行建模。 Hide-and-Seek [25] 在训练时隐藏随机帧序列，以强制网络响应多个相关部分。 然而，随机隐藏帧并不总能保证新部分的发现，还会扰乱训练过程。 最近，钟等人。 [52] 通过从输入视频中删除前任分类器的预测，迭代地训练一系列分类器以找到互补的片段。 这种方法的主要缺点是训练多个分类器需要额外的时间成本和计算费用。 动作-上下文分离的另一个挑战本质上是棘手的，并且在文献中仍未探索。 UntrimmedNet [45] 中的选择模块旨在消除不相关的背景剪辑，而不是语义相关的上下文。研究人员还研究了具有其他类型的弱监督的动作定位，例如电影脚本 [13]、有序动作列表 [11] 和网络图像 [14]。

Diversity Loss 最初是为文本嵌入 [32] 引入的，以提取句子的不同方面。 最近，李等人。 [29]利用多样性损失来处理人员重新识别中的遮挡。 与以前的工作不同，我们使用多样性损失来建模动作完整性，它具有不同的规范和动机。

## 3. 提议的方法(Proposed Method)

在本节中，我们提出了用于弱监督时间动作定位的方法。 输入是一个未修剪的视频，具有不同的帧长度。 设一个one-hot向量表示ground truth video-level类别标签，其中是动作类的数量，表示新添加的背景类。 在测试时间内，每个测试视频的输出是一组局部动作实例 ，其中 和 表示第 i 次检测的开始时间和结束时间，表示预测类别， 表示置信度分数。

### 3.1 硬负视频生成(Hard Negative Video Generation)

弱监督模型倾向于将真实动作与其周围的上下文（即硬否定）混淆，尤其是当上下文出现在该类的大多数视频中时。 我们观察到，正是运动使动作与其上下文不同。 动作必须涉及人类或其他主体的移动，而上下文剪辑被允许保持静态（例如，静态台球桌）。 因此，我们使用固定视频剪辑生成硬负训练数据，并用新的背景类标记它们。 具体来说，对于训练集中的每个视频，我们使用 TV-L1 算法 [50] 计算其光流，并对每一帧的强度进行平均。 由于动作类别之间的运动幅度不同，甚至某些动作表现出轻微的运动，因此从每个视频中单独挑选出具有最低光流强度的一小部分预定义视频帧。 然后将从同一视频中挑选的帧连接成一个伪视频，该视频用背景类标记并添加到训练集中。 我们期望生成的视频部分包含所提出的网络的硬底片和丢弃提示，以应对动作-上下文分离的挑战。 补充材料中提供了详细信息和生成的视频示例。

### 3.2 多分支网络

为了模拟动作完整性，设计了一个多分支网络，使得每个分支关注不同的动作部分。如图2所示，所提出的多分支网络由特征提取模块、嵌入模块、多分支分类模块和时间注意模块组成，具体如下

**特征提取模块**。给定一个输入视频，首先通过预训练的深度网络提取片段特征序列，其中T表示片段数，D表示特征维数。提取的特征序列提供了输入视频的外观和运动的高级表示，并被送入网络的下一层。注意，T和D取决于特征提取网络的选择。在实验中，我们关注两个现成的模型，即UntrimmedNet[45]和I3D [7]。

**嵌入模块**。特征提取模块之后是嵌入模块。由于这些特征最初可能不是为弱监督的动作定位而训练的，因此需要特定于任务的特征嵌入。我们利用时间卷积层和ReLU激活层来嵌入特征：

(1)

其中表示卷积运算，和是时间滤波器的权重和偏差， 表示学习的嵌入，F是滤波器的数量。时间卷积集成了来自相邻时间位置的信息，使网络能够捕获时间结构。然后将嵌入的特征序列传递给后续层。

**多分支分类模块**。在这个模块中，K个分类分支并行组织，以发现动作的互补部分。每个分支将嵌入的特征序列输入到时间卷积层，并输出分类分数序列(a sequence of classification scores)：

(2)

其中，、和分别是第k分支中的分类分数、过滤器权重和过滤器偏差。然后，每个沿着类别维度通过softmax，在每个时间位置产生类别分布：

其中被称为类激活序列（CAS）。为了清楚起见，我们在本文中使用条形符号表示它已经经历了softmax。对于动作完整性建模，我们期望来自多个分支的案例彼此不同。然而，如果没有约束，分支可以懒洋洋地专注于同一个动作部分。为了避免分支给出相同结果的退化情况，对这些情况施加基于余弦相似性的多样性损失：

(4)

这是来自每两个分支的 CAS 之间的余弦相似度，是所有分支对和动作类别的平均值。 表示来自第 i 个分支的 c 类的激活序列， 是归一化因子。 通过最小化这种多样性损失，鼓励分支在不同的动作部分上产生激活。 然后来自多个分支的 CAS 被平均并沿类别维度通过 softmax：

(6)

其中 被称为平均 CAS，它结合了所有部分激活并编码完整动作。

此外，当背景类的分数很大时，softmax 操作会抑制动作类的激活，从而减少上下文剪辑的误报

我们凭经验注意到，来自某些分支的 往往几乎全为零，而来自其他分支的 会爆炸，这可能会破坏训练过程。 更重要的是，如果一个分支占主导地位，则平均 CAS 会有效地响应单个动作部分而不是整个动作。 从另一个角度来看，这些平行的分支可以被认为是一种对抗关系，相互竞争以寻找不同的判别动作片段。 预计分支平衡以具有可比强度。 在生成对抗网络[18]的训练策略中可以看到类似的想法。因此，我们在没有 softmax 的原始分数序列的范数上引入了另一个正则化项：

这是与 规范的偏差，在分支和类别上平均。 配备多样性损失和范数正则化，多分支设计能够在没有完全监督的情况下发现不同的动作部分，从而对动作完整性进行建模。

**时间注意模块(**Temporal attention module)。 由于输入视频未经修剪并且包含不相关的背景，我们利用时间注意模块来学习视频片段的重要性。 注意模块将嵌入的特征序列馈送到时间卷积层，然后是沿着时间维度的 softmax：

(8)

其中 和 是时间滤波器的权重参数和偏差，表示学习到的与类别无关的注意力序列。 为了获得视频级别的分类预测，我们对由注意力加权的 的总和沿类别维度执行 softmax：

(9)

其中 是动作类的概率分布，包括背景类。 然后计算其与基本事实的交叉熵，即标准 MIL 损失：

(10)

最后，我们将 MIL 损失与多样性损失和范数正则化结合起来：

(11)

其中和 是系数。 所有三个分量都至少有子梯度，并且可以使用梯度下降来最小化。

### 3.3 动作定位(Action Localization)

在测试期间，我们利用经过训练的多分支网络对测试视频进行分类并定位动作。由于一个视频中可能会发生多种动作，我们首先对视频级分类分数设置阈值。给定一个测试视频，我们检测每个非背景类别的动作实例，其中 大于 0.1。 然后我们对 类的平均 CAS 进行阈值化，即 \*，以定位动作实例。 令 表示相应的输出检测。 类似于 [38] 中提出的 Outer-InnerContrastive 损失，我们使用实例本身的平均激活与其周围区域之间的对比对每个局部实例进行评分：

(12)

其中 [.] 表示连接， 是膨胀长度。 视频级分数也与系数 相结合

## 4. 实验

在本节中，我们首先讨论数据集和我们的实现细节。 然后提出了所提出的方法与最先进的方法之间的比较。最后，我们通过消融研究检查每个模型组件的影响。 在补充材料中，报告了更多的实验结果。

### 4.1 数据集

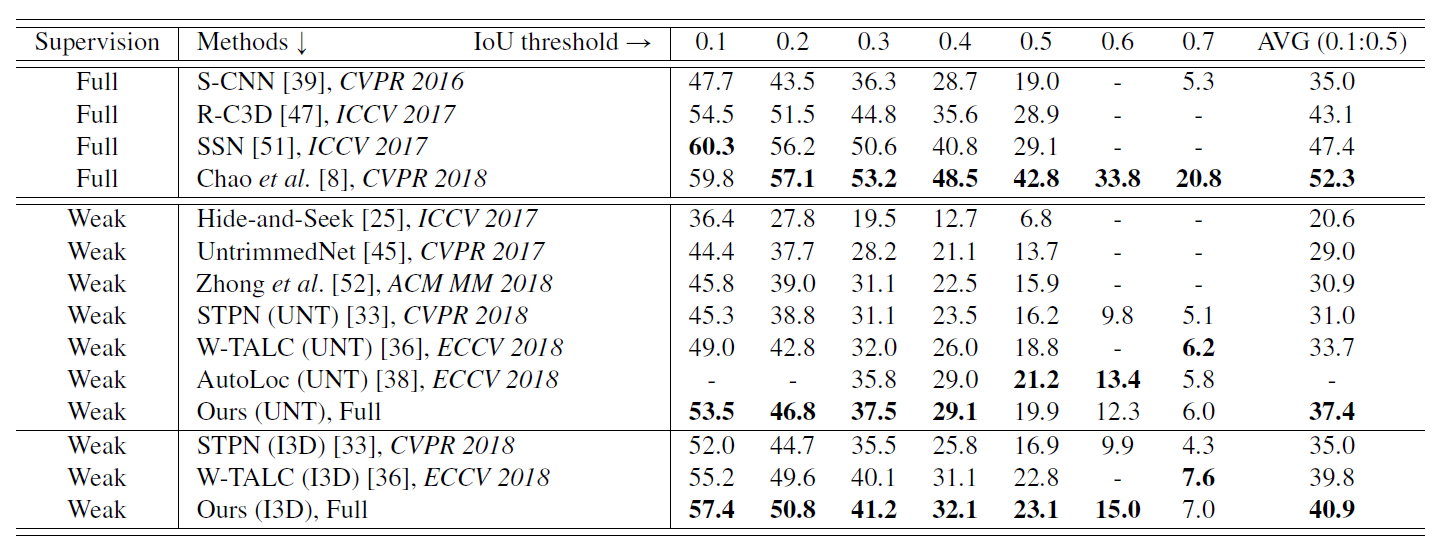
在两个大型基准上进行了广泛的实验：THUMOS'14 [21] 和 ActivityNet [6]。两个数据集中的视频均未修剪，仅使用视频级别的类别标签进行训练。

**THUMOS’14**。 THUMOS'14 的一个子集包括 20 个动作类，带有时间注释并用于定位任务。 按照之前的惯例，我们使用 200 个视频的验证集进行训练，使用 213 个视频的测试集进行评估。 从训练数据中，生成了 152 个硬否定视频。 这个数据集每个视频都有大量的动作实例，视频的长度变化很大。

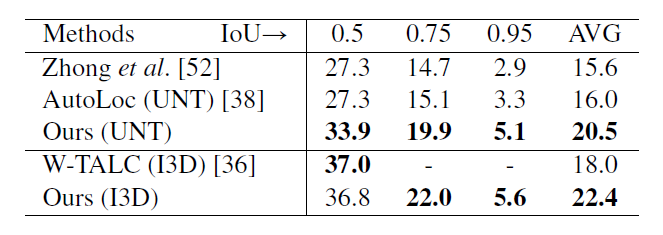
**ActivityNet**。 在 ActivityNet 的两个发布版本上进行了实验。 ActivityNet1.3 涵盖 200 个动作类，由 10,024 个训练视频、4,926 个验证视频和 5,044 个测试视频组成，其中使用训练视频生成了 7323 个硬否定视频。 我们在训练集上进行训练，并在验证集和测试集上报告结果。 为了便于比较，我们还评估了 ActivityNet1.2，它是 1.3 版的一个子集，它有 4,819 个训练视频、2,383 个验证视频、2,480 个测试视频和 100 个类。 在 ActivityNet1.2 上生成了 3469 个硬否定视频。 我们使用训练集进行训练，使用验证集进行评估。

**Evaluation Metrics评估指标**。 我们遵循标准评估协议并报告在时间交叉联合（IoU）的不同阈值下的平均平均精度（mAP）。 使用数据集提供的评估代码计算 mAP 值。 THUMOS'14 上的所有结果都是三次运行的平均值。 ActivityNet1.3 测试集上的性能是通过将结果提交给评估服务器获得的

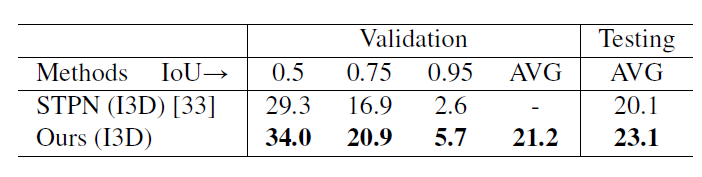
### 4.2 实现细节(Implementation Details)



***Table1: THUMOS'14 测试集的结果。 报告了不同 IoU 阈值下的 mAP 值，AVG 列表示 IoU 阈值从 0.1 到 0.5 的平均 mAP。 UNT 和 I3D 分别是 UntrimmedNet features 和 I3D features 的缩写。凭借 UntrimmedNet 和 I3D 功能，我们的完整模型在大多数 IoU 上都优于最先进的方法***



***Table2: ActivityNet1.2 验证集的结果。 AVG 列表示 IoU 阈值 0.5:0.05:0.95 处的平均 mAP。所提出的方法大大超过了以前的方法。***

******

***Table3: ActivityNet1.3 上的结果。 AVG 列表示 IoU 阈值 0.5:0.05:0.95 处的平均 mAP。 我们的方法也实现了卓越的性能***

尝试了两个具有双流架构的深度网络进行特征提取，即 UntrimmedNet [45] 和 I3D [7]，它们在训练期间进行了预训练和固定。UntrimmedNet 在 ImageNet [10] 上进行了预训练，并将 1 个 RGB 帧和 5 个堆叠的光流帧的视频片段作为输入。 I3D 在 Kinetics [7] 上进行了预训练，并将不重叠的 16 帧块作为两个流的输入。 UntrimmedNet 的视频片段每 15 帧采样一次，I3D 的视频片段每 16 帧采样一次。 两个网络的每个流中的特征维度都是 1024。我们对 UntrimmedNet 采用 RGB 和光流流的早期融合，对 I3D 采用晚期融合。

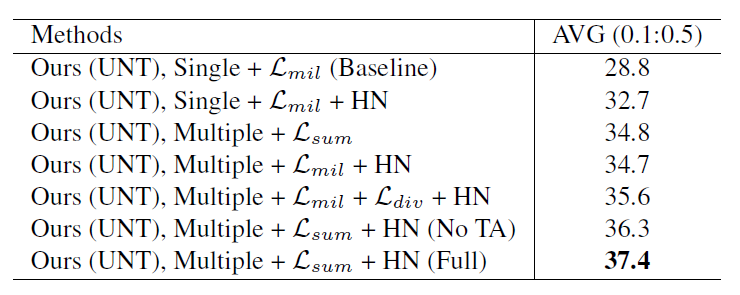
所提出的方法是用 PyTorch [35] 实现的。网络参数是使用带有 Adam 优化器的小批量随机梯度下降 [24] 来学习的。 在多分支分类模块中，分支编号设置为。时间卷积的核大小在分类模块中设置为 3，在嵌入模块和注意模块中设置为 1。 嵌入特征的维数设置为。Eq.(11)中的系数和均设置为 0.2，Eq.(12)中设置为 0.25。 硬负挖掘(hard negative mining)的选择率选择为 25%。 补充材料中提供了其他详细信息。

### 4.3 与 State-of-the-art 的比较

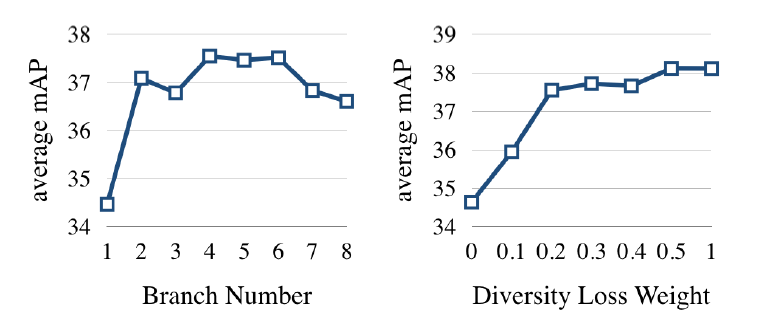
THUMOS'14 测试集的实验结果如表 1 所示。我们提出的多分支网络以及硬负挖掘与现有的弱监督时间动作定位方法以及几个完全监督的方法进行了比较。 无论选择何种特征提取网络，我们的模型在大多数 IoU 阈值上都优于以前的弱监督方法。 由于观察到我们的模型有时会产生导致误报的过于完整的实例，因此在 IoU 较高时增益并不那么明显。 请注意，AutoLoc [38] 回归了定位的时间动作边界，因此在更高的 IoU 阈值下获得了高 mAP，而我们只是在 CAS 上设置阈值并仍然获得可比较的结果。 我们认为，如果将他们的方法和我们的方法结合起来，可以进一步提升性能。

表 2 给出了 ActivityNet1.2 验证集的结果，而 ActivityNet1.3 的验证集和测试集的结果在表 3 中报告。在这个大型数据集的两个版本上，所提出的方法都显着优于 state-of-the-art，验证了处理动作完整性建模和上下文分离的有效性。

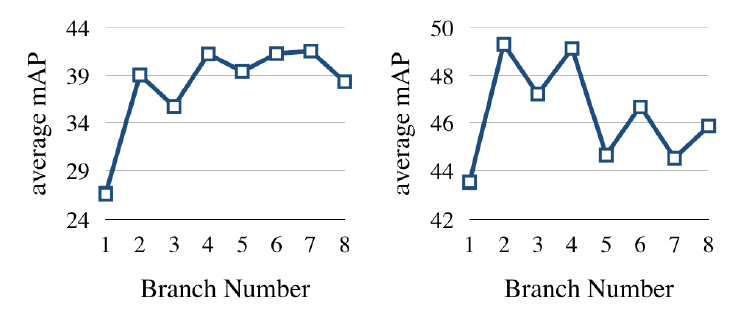
### 4.4 消融研究(Ablation Studies)



***Table 4 THUMOS'14 测试集的消融研究结果。 “Single”和“Multiple”表示分类模块中的分支数，“HN”表示训练时使用硬负视频。 “TA”表示时间注意模块。***

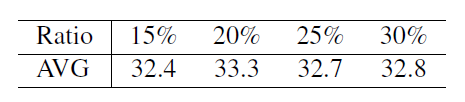


***Figure3 : 左：分支编号的实验。 右：关于多样性损失权重的实验。 报告 IoU 阈值从 0.1 到 0.5 的平均 mAP***



***Figure4: 具有不同分支号的特定于类的结果。 最佳分支数取决于动作的复杂性。***

***左：Shotput 的结果。右：Cliff Diving的结果。 报告 IoU 阈值从 0.1 到 0.5 的平均 mAP***



***硬负挖掘中选择率的影响。AVG 表示 IoU 阈值从 0.1 到 0.5 的平均 mAP。***

为了分析每个模型组件的贡献，我们进行了一组消融研究，THUMOS 的 14 个测试集的结果如表 4 所示。我们的最佳模型与基线和其他配置进行了比较，并删除了以下每个组件：1 ) 多分支设计 2) 硬负生成 3) 多样性损失和范数正则化 4) 只有范数正则化 5) 时间注意模块。 结果表明，所有这些组件都需要达到最佳性能，而多分支设计尤为重要。 此外，我们对 UntrimmedNet 特征进行了实验，以研究分支数、多样性损失权重以及选择率在硬负挖掘中的影响。

**Branch Number**。 对Branch Number进行对比实验，Branch Number K由1改为8。 为了避免在测试数据上调整参数，我们在 THUMOS 的 14 验证集（即用于训练的验证集）上进行了此实验。 结果如图 3 所示，多分支模型明显超过单分支模型，二到八分支模型之间差异不显着。由于动作的复杂性因类别而异，因此每个动作类别的最佳分支数可能不同。 如图 4 中给出的示例，Shotput 等复杂动作由更多部分组成，因此需要更大的分支数，而 Cliff Diving 等简单动作的结构只需两个分支即可捕获。

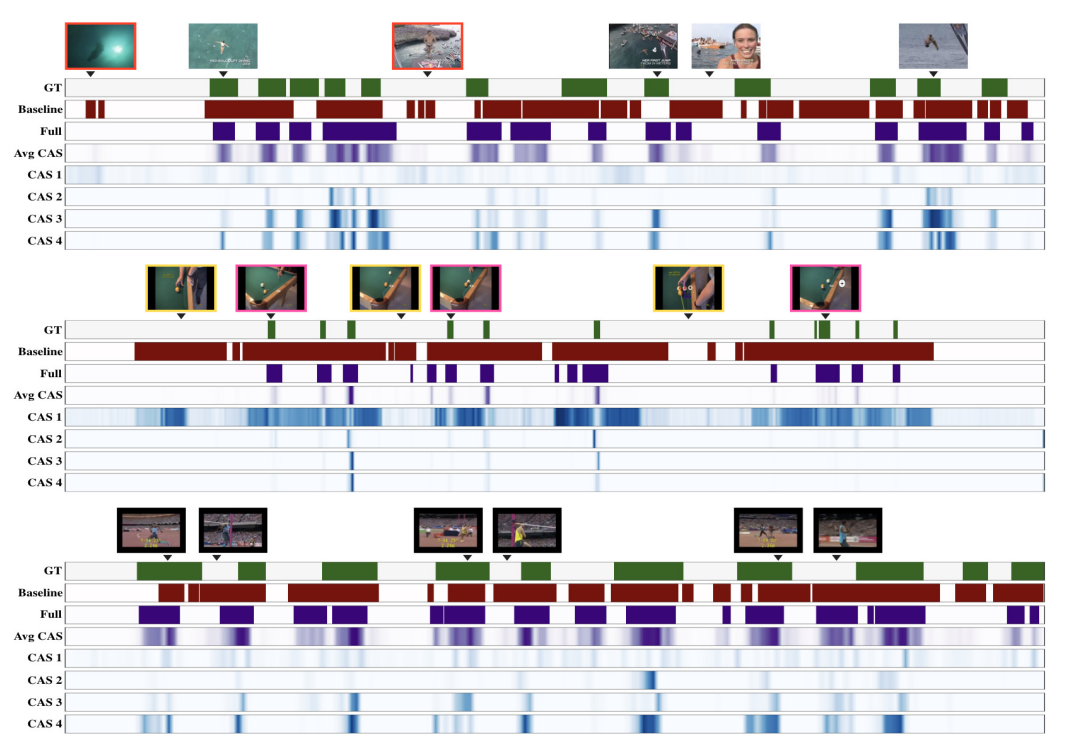
**多样性损失的权重**。 另一个关于多样性损失权重的比较实验是在 THUMOS 的 14 个验证集上进行的，结果如图 3 所示。由于范数正则化和多样性损失旨在联合约束多个分支，因此系数和设置为相同的变化 价值。 实验表明，我们的模型在大于 0.2 时对多样性损失权重不敏感，证明了该方法的鲁棒性。

**硬负挖掘中的选择比**。 选择静态帧时尝试了几种不同的百分比。 由于硬否定视频是从验证集生成的，因此该实验是在 THUMOS 的 14 个测试集上进行的。 单分支模型产生的结果如表 5 所示，在不同比率下保持稳定

### 4.5 定性结果(Qualitative Results)

我们在图5中绘制了几个有趣的局部行动示例和相应案例，以显示定性应对这两个挑战的有效性。示例来自THUMOS'14测试集，使用未经修剪的网络功能。在第一个潜水示例中，显示在红色边界框中的不完整动作，例如仅进入水部分的动作和仅站在平台部分的动作，由单个分支发现，但由于不完整性而排除在最终预测之外。在台球的第二个例子中，语义上与真实动作（粉色框）相似的评论片段（黄色框）通过硬负面生成有效地过滤掉。在跳高的第三个例子中，来自多个分支的案例非常不同，将不同的动作部位定位。

## 5 讨论和未来工作



***Figure5： 三个预测例子。每种情况下的八个条形码分别是：1）行动的基本真相实例2）基线模型的预测，即没有硬负生成的单分支网络3）我们完整模型的预测4）我们完整模型的平均CA 5-8）来自四个分支的案例。上图：跳水。不完整操作（红色框）仅激活单个分支，因此被平均CA排除。中间：台球。该方法关注真实动作（粉色框），减少语义相似上下文（黄色框）的误报。垫底：跳高。在所有示例中，每个分支都输出一个不同的CA。***

我们设计了一个简单而有效的数据生成方案来分离行动背景，而背后的假设可能并不适用于所有情况。我们发现它的效果与动作类密切相关，细节见补充部分。未来，更先进的技术，如生成性对抗网络，可以被应用于更深入地挖掘硬的负面影响。对于动作完备性建模，提出的多分支模块以无监督的方式自动发现不同的动作部分。在实践中，学习到的动作部分可能并不完全对应于语义上有意义的子动作。相反，该模型可以捕捉不同的动作模式、方面、阶段或其他底层结构，这取决于哪种表现形式对学习目标最有利。按照将一个动作分成几个部分的关键思想，弱监督的时间动作定位可能有许多潜在的未来方向，包括但不限于1）使用学习的部件表示来理解动作或测量动作复杂性2）建模动作部件的时间配置3）分层表示动作以处理歧义或主观注释偏差。

## 6 结论

在这项工作中，我们发现了对时间动作定位的弱监督带来的两个挑战，即动作完整性建模和动作上下文分离。为了应对第一个挑战，提出了一个多分支网络来寻找不同的动作部分，从而在它们的完整性中定位动作实例。同时，我们挖掘了硬否定来处理第二个问题，即动作-语境分离。在两个基准测试上的实验表明，我们的框架有效地解决了这两个问题，并且优于最先进的方法

**致谢**。中国国家自然科学基金项目（2015CB351803）和国家自然科学基金（中国自然科学基金）61572042, 61527804, 61625201年度合同（973计划）部分资助了这项工作。我们感谢北京大学提供计算资源的高性能计算平台。感谢Jingjia Huang提供的光流数据。

引用：

[1] Jake K Aggarwal and Michael S Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011.

[2] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting actions in videos and its application to temporal action localization. In *The European* *Conference on Computer Vision (ECCV)*, September 2018.

[3] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, V´ıctor Ponce-L´opez, Xavier Bar´o, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. A survey on deep learning based approaches for action and gesture recognition in image sequences. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International* *Conference on*, pages 476–483. IEEE, 2017.

[4] S Buch, V Escorcia, B Ghanem, L Fei-Fei, and JC Niebles. End-to-end, single-stream temporal action detection

in untrimmed videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2017.

[5] Fabian Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. SCC: Semantic context cascade for efficient action detection. In *The IEEE Conference on Computer* *Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *The IEEE* *Conference on Computer Vision and Pattern Recognition* *(CVPR)*, June 2015.

[7] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *The* *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[8] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. Rethinking the Faster R-CNN architecture for temporal action localization. In *The IEEE Conference on Computer Vision* *and Pattern Recognition (CVPR)*, June 2018.

[9] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S. Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *The IEEE International Conference* *on Computer Vision (ICCV)*, Oct 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image

database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[11] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *The* *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *The IEEE* *Conference on Computer Vision and Pattern Recognition* *(CVPR)*, June 2015.

[13] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *The IEEE International Conference on Computer* *Vision (ICCV)*, pages 1491–1498. IEEE, 2009.

[14] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *The European Conference on Computer Vision (ECCV)*, pages 849–866. Springer, 2016.

[15] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Cascaded boundary regression for temporal action detection. In *Proceedings of the British Machine Vision Conference (BMVC)*,2017.

[16] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi,Balakrishnan Varadarajan, Henry C Lin, Lingling Tao,Luca Zappella, Benjamın B´ejar, David D Yuh, et al. JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *MICCAI Workshop: M2CAI*, volume 3, page 3, 2014.

[17] Georgia Gkioxari and Jitendra Malik. Finding action tubes.In *The IEEE Conference on Computer Vision and Pattern* *Recognition (CVPR)*, June 2015.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, DavidWarde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in* *Neural Information Processing Systems (NIPS)*, pages 2672–2680, 2014.

[19] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. Going deeper into action recognition: A survey. *Image and Vision* *Computing*, 60:4–21, 2017.

[20] Rui Hou, Rahul Sukthankar, and Mubarak Shah. Real-time temporal action localization in untrimmed videos by subaction discovery. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 2, page 7, 2017.

[21] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban,Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos “in the wild”. *Computer Vision and Image Understanding*, 155:1–23, 2017.

[22] Soo Min Kang and Richard P Wildes. Review of action recognition and detection methods. *arXiv preprint*

*arXiv:1610.06906*, 2016.

[23] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *The* *IEEE Conference on Computer Vision and Pattern Recog-* *nition (CVPR)*, June 2014.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on* *Learning Representations (ICLR)*, 2015.

[25] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *The IEEE International* *Conference on Computer Vision (ICCV)*, Oct 2017.

[26] Ivan Laptev. On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123, 2005.

[27] Colin Lea, Michael D. Flynn, Rene Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *The IEEE Conference* *on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[28] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *The IEEE Conference on Computer Vision* *and Pattern Recognition (CVPR)*, pages 1346–1353. IEEE, 2012.

[29] Shuang Li, Slawomir Bak, Peter Carr, and Xiaogang Wang. Diversity regularized spatiotemporal attention for videobased person re-identification. In *The IEEE Conference* *on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[30] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 2017 ACM on* *Multimedia Conference*, pages 988–996. ACM, 2017.

[31] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *The European Conference on* *Computer Vision (ECCV)*, September 2018.

[32] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *arXiv preprintarXiv:1703.03130*, 2017.

[33] Phuc Nguyen, Ting Liu, Gautam Prasad, and Bohyung Han. Weakly supervised action localization by sparse temporal pooling network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[34] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. Action and event recognition with fisher vectors on a compact feature set. In *The IEEE International Conference on Computer* *Vision (ICCV)*, December 2013.

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.

[36] Sujoy Paul, Sourya Roy, and Amit K. Roy-Chowdhury. WTALC: Weakly-supervised temporal activity localization and classification. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[37] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. CDC: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[38] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. AutoLoc: Weakly-supervised temporal action localization in untrimmed videos. In *The European* *Conference on Computer Vision (ECCV)*, September 2018.

[39] Zheng Shou, DongangWang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *The IEEE Conference on Computer Vision and Pattern* *Recognition (CVPR)*, June 2016.

[40] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NIPS)*, pages 568–576, 2014.

[41] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[42] Sarvesh Vishwakarma and Anupam Agrawal. A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10):983–1009, 2013.

[43] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *The IEEE International Conference* *on Computer Vision (ICCV)*, December 2013.

[44] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification.

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[45] LiminWang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *The IEEE Conference on Computer Vision* *and Pattern Recognition (CVPR)*, July 2017.

[46] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *The European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2016.

[47] Huijuan Xu, Abir Das, and Kate Saenko. R-C3D: Region convolutional 3D network for temporal activity detection. In *The IEEE International Conference on Computer Vision* *(ICCV)*, Oct 2017.

[48] Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei- Fei. End-to-end learning of action detection from frame glimpses in videos. In *The IEEE Conference on Computer* *Vision and Pattern Recognition (CVPR)*, June 2016.

[49] Zehuan Yuan, Jonathan C. Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *The IEEE Conference on Computer Vision and Pattern* *Recognition (CVPR)*, July 2017.

[50] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L1 optical flow. In *Joint* *Pattern Recognition Symposium*, pages 214–223. Springer, 2007.

[51] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *The IEEE International* *Conference on Computer Vision (ICCV)*, Oct 2017.

[52] Jia-Xing Zhong, Nannan Li, Weijie Kong, Tao Zhang, Thomas H Li, and Ge Li. Step-by-step erasion, one-by-one

collection: A weakly supervised temporal action detector. In *Proceedings of the 2018 ACM on Multimedia Conference*. ACM, 2018.

[53] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[54] Zhi-Hua Zhou. Multi-instance learning: A survey. *Department of Computer Science & Technology, Nanjing University, Tech. Rep*, 2004.