

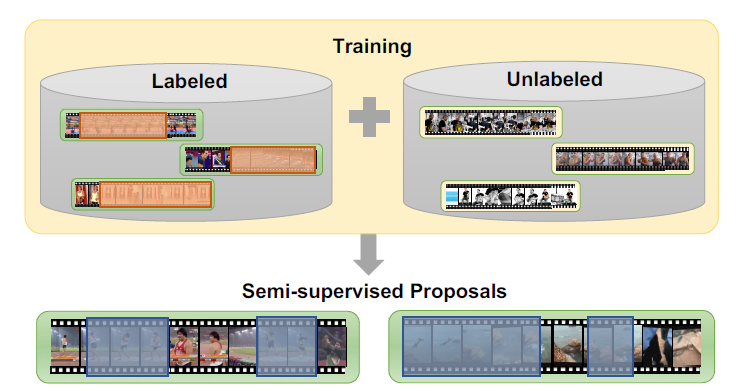


**使用较少的标签学习临时行动建议(**Temporal Action Proposals)

## 摘要

时间行动建议是当今行动检测管道中的一个常见模块。 大多数当前训练动作提议模块的方法都依赖于完全监督的方法，这些方法需要在长视频序列中使用大量带注释的时间动作间隔。 这需要大量的注释成本和努力促使我们研究在较少监督的情况下训练提案模块的问题。 在这项工作中，我们提出了一种专门为训练时间动作提议网络而设计的半监督学习算法。 当只有少量标签可用时，我们的半监督方法比完全监督的对应物和其他强大的半监督基线生成的提议要好得多。 我们在两个具有挑战性的动作检测视频数据集 ActivityNet v1.3 和 THUMOS14 上验证了我们的方法。 我们表明，我们的半监督方法始终匹配或优于完全监督的最先进方法。

## 1 介绍



***Figure1：我们的半监督框架只有部分训练视频标有基本事实建议，可以生成比最先进的全监督方法质量更好的时间动作建议***

世界上有数以百万计的摄像机，每天都会产生和传输大量的视频。这些视频中的一个非常重要的主题是人类的活动和行为。这促使计算机视觉界研究从视频采集中理解动作的算法。动作理解的一项重要任务是动作检测，即时间动作定位，其目标是在长视频序列中暂时定位所有感兴趣的动作。解决这个问题的一种常见方法是首先生成时间动作建议，以定位感兴趣的时间间隔，然后将其输入分类器以获得相应的动作标签。在本文中，我们主要关注时间动作提议模块。

为了达到较高的预测精度，大多数现有的最先进的时间动作建议算法都使用有监督的深度学习方法[3,14,15,23]。这种方法需要大量有标签的视频。与图像识别等其他视觉任务中的标记不同，在未剪辑的视频中标记动作的时间边界要耗时得多。另一方面是无监督学习方法[34]，培训不需要标签。虽然它们没有标签的负担，但在许多任务中的总体性能通常不可避免地低于监督方法。

当有大量数据可用但只有一小部分被标记时，半监督学习是一种非常适合的解决方案。与无监督学习不同，半监督学习仍然利用标记数据作为高预测精度的强监督。与监督学习相比，半监督学习可以利用未标记的数据，因此不太可能在小的标记数据集上过度拟合。半监督学习在图像分类中是有效的[21,25,29,36]，但从未被探索用于帮助生成临时行动建议。在我们的问题设置（见图1）中，我们假设在训练期间，只有一部分视频带有用于监督学习的动作的时间边界标签。与此同时，培训过程中还可以利用其他没有标签或注释的视频。通过将从标记集提取的知识扩展到未标记集，我们可以获得一个更健壮的模型，因为未标记数据可以发挥正则化作用。

半监督学习方法背后的一个核心理念是训练模型，使其具有平滑一致的分类边界，对随机扰动具有鲁棒性。为了找到一个平滑的数据流形，Tarvainen等人[36]提出了Mean-Teacher，将不同训练迭代中的“学生”模型平均化为“教师”模型。我们在模型设计中采用了这种架构。为了提高模型的鲁棒性，在学生模型的输入上引入随机扰动至关重要。特别是对于视频中的时间动作建议任务，扰动的设计应该有利于序列学习。然而，之前的工作并没有对视频等序列数据提出适当的扰动。

我们提出了两种类型的连续扰动：时间扭曲和时间掩蔽。时间扭曲是一种重采样层，它沿时间维度扭曲视频序列，为时间敏感任务（如时间动作建议）提供扰动。时间屏蔽随机屏蔽输入视频的某些帧。在培训期间，蒙面学生模型只看到部分视频，而他们被鼓励预测与无障碍教师模型预测相同的边界。这些连续的扰动使我们的优化模型更加稳健，并能更好地推广到看不见的数据。

我们的主要贡献如下：（1）据我们所知，我们是第一个将半监督学习纳入时间行动方案以实现标签效率的人。（2） 我们为这个半监督框架设计了两种基本类型的顺序扰动，并在时间动作建议的关键实验中根据强半监督基线对它们进行了验证

## 2 相关工作

**时间行为检测和建议**。给定一段长且未经剪辑的视频，时间动作检测旨在定位每个动作实例及其开始和结束时间以及动作类[4、12、14、16、22、33、40]。

传统上，许多方法通过以滑动窗口方式彻底应用动作分类器来解决这个问题[13、19、26、27、37、39]。这些方法在计算成本方面通常效率低下，因为它们需要覆盖整个未剪辑视频中每个位置不同长度的时间窗口。

受最近图像目标检测的proposal plus分类方法的成功启发，另一组两阶段方法首先在视频中提出动作无关的时间段，然后对修剪片段的动作进行分类。Buch等人[3]提出了一种执行单流时间动作建议生成的网络，避免了滑动窗口带来的计算成本。Shou等人[32]使用3D ConvNet生成临时提案。还有一些端到端的框架，可以实现提案生成和行动分类的联合优化。Buch等人。[2] 为端到端时间动作定位课程培训引入语义约束。Chao等人[8]采用更快的R-CNN[30]进行动作定位任务。

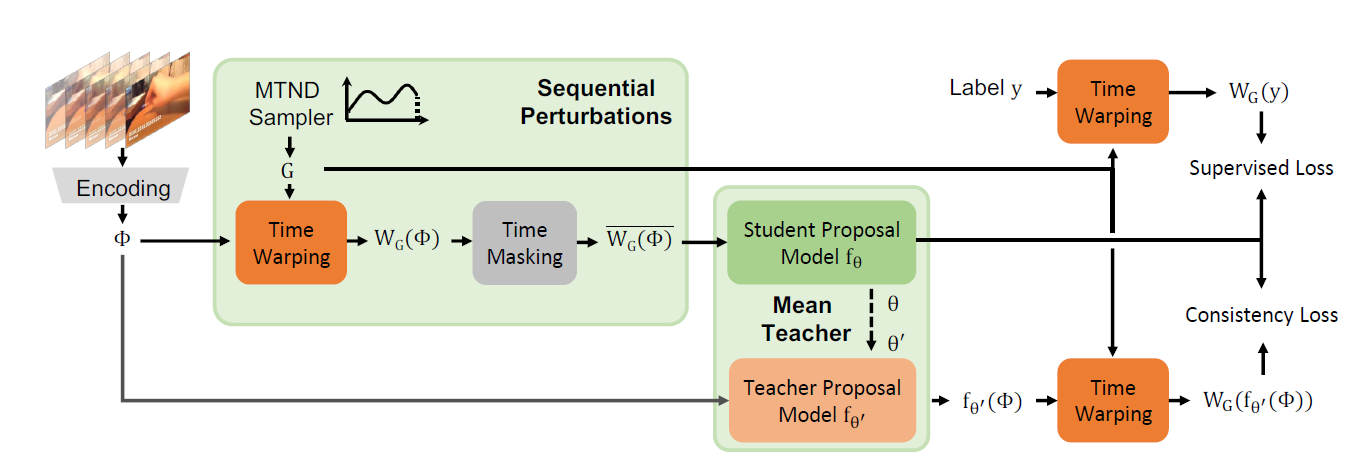
在上述方法中生成的建议通常依赖于预定义的锚，缺乏时间界限的灵活性和准确性。相反，Zhao等人[41]将提案生成问题简化为对每个短视频片段的活动性进行分类，并通过分水岭算法进行后处理。Gao等人[15]和边界敏感网络（BSN）[23]进一步推断视频片段是获取更精确边界的行动的开始还是结束，其中BSN已成为ActivityNet Challenge[5]上临时行动建议任务的最新技术。

之前的研究致力于开发更好的行动建议模型，并通过标记视频进行训练。同时，我们探索如何利用未标记的视频进一步提高提议和检测性能。在这项工作中，由于BSN的优越性能，我们将重点评估我们的半监督框架，尽管我们的框架的灵活性也允许它与其他时间动作建议架构相结合。

**半监督深度学习**。半监督学习有着跨越数十年的丰富历史[9,42]。我们的重点不是全面的回顾，而是半监督的深度学习。一种常见的方法是通过联合优化标记数据的监督分类损失和标记和未标记数据的额外无监督损失来训练神经网络[21,25,29,36]。

一致性正则化被广泛应用于无监督损失，它鼓励模型在原始输入或中间特征映射受到扰动时生成一致的输出。

这里我们总结了一些使用一致性正则化的半监督深度学习的例子。梯形网络[29]将重建分支合并为无监督任务；在每一个训练步骤中，它们都会在编码和解码的激活图之间造成一致性损失。模型[21]简化了梯形网络，只会在数据受到不同扰动的输出之间造成一致性损失。接下来，时间加密[21]将一致性损失应用于模型输出和更稳定的目标：每个历元模型输出的指数移动平均值。与平均输出不同，更强大的平均教师[36]将每个训练步骤（又称“学生”模型）的模型权重平均化为一个单独的“教师”模型，其输出作为一致性损失的目标。与上述方法正交，虚拟对抗训练（VAT）[25]提出使用虚拟对抗噪声代替随机噪声作为数据扰动。在我们的工作中，我们还对学生和教师模型的输出进行了一致性正则化，并提出了时间扭曲和时间掩蔽作为视频数据的数据扰动。



***Figure2：我们的方法概述。给定一个未经剪辑的视频作为输入，我们首先将其编码为一个特征序列。接下来，对应用包括时间扭曲和时间掩蔽在内的序列扰动，学生提案模型将该扰动序列作为输入。相反，教师模型直接预测无障碍。最后，对学生模型进行了联合优化，对标记的视频应用监督损失，对所有视频应用一致性损失***

半监督学习也被应用到序列学习中。Dai等人[11]提出了一种用于文本分类的序列自动编码器。Pr’emont Schwarz等人[28]将梯形网络与递归神经网络相结合，并在被遮挡的移动MNIST数据集上评估其图像分类模型。Clark等人[10]提出了针对多种语言任务的交叉视角训练。Miyato等人[24]将VAT[25]应用于文本分类。虽然不是为视频分析而设计的，但上面的一些方法[10,28]也包含了在图像中的补丁或句子中的单词上进行掩蔽的想法，它们启发了我们的时间掩蔽。

也有关于弱监督学习的工作，用于时间动作检测[1,7,17,31]，这与我们的半监督设置不同。在弱监督的时间动作检测中，部分训练数据被完全标记为时间边界和动作类，而其余数据则被标记为“弱”标签，即视频级别的类或视频中动作的顺序列表。相反，我们不假设在我们的半监督培训中使用的未标记视频有任何类型的标签，这需要一个更难但更有效的标签任务。

## 3 技术方法(Technical Approach)

我们的主要目标是用相对较少的标签生成高质量的临时行动建议。这就要求我们通过一个强大的有监督的提议模型来最好地利用标记的数据，同时，通过一个设计用于视频理解的无监督辅助任务来利用未标记的数据。虽然我们的方法对特定的提案方法不可知，但为了验证半监督框架，我们在最先进的完全监督提案生成网络——边界敏感网络——的基础上构建了我们的模型[23]。我们用两种类型的顺序扰动扩展了平均教师框架[36]，用于训练提议模型：时间扭曲和时间掩蔽。请参见图2，作为我们方法的概述。

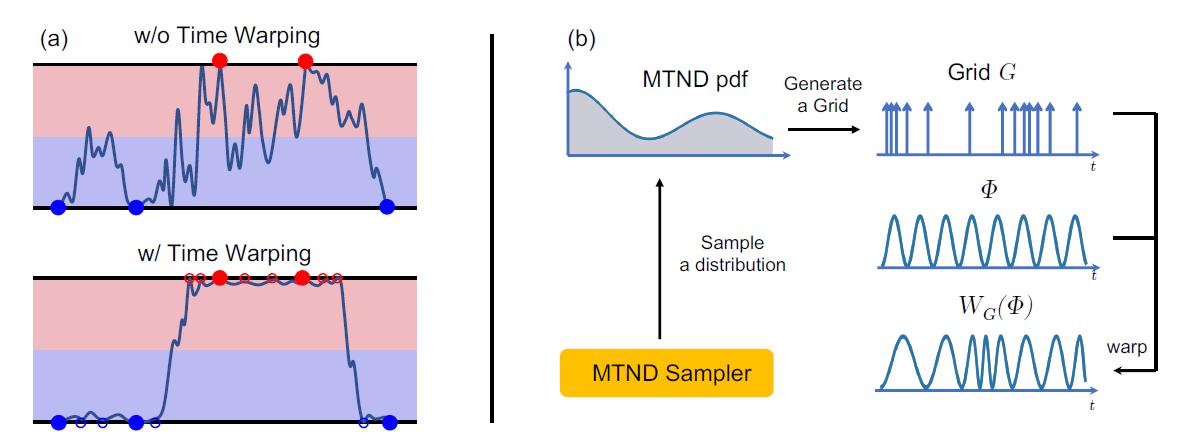
### 3.1 视频编码

视频编码的目的是获得压缩视频表示，它捕获视频的外观和运动模式。给定一个以N帧为输入的未剪辑视频，我们首先将其分成不重叠的短片段，每个片段包含帧，形成一系列片段，其中。如之前的工作[6,38]所示，外观和运动特征都有助于理解动作，因此我们对RGB帧和每个视频的光流进行编码，然后将编码的向量连接起来。特别是，我们使用[38]作为视频编码器，就像在完全监督的基线[23]中一样。编码器生成一系列特征向量 .。然后我们将特征向量序列以小批量的形式输入以下模块。标记和未标记的视频共享同一个视频编码器，并且它们在同一个小批量中共存。

### 3.2 临时行动建议模型

我们的半监督模型具有足够的灵活性，可以建立在各种完全监督的时间动作建议网络上，只要它们以序列数据为输入。具体来说，我们选择了边界敏感网络（BSN）[23]，它是2018年ActivityNet挑战赛中临时行动建议任务中的佼佼者。

第一步执行与[23]中相同的视频编码，然后将直接输入BSN提案模型。BSN由两个可训练模块组成：临时评估模块（TEM）和提案评估模块（PEM）。视频编码后，TEM将片段特征序列作为输入。序列 通过时间卷积层生成三系列概率信号：动作性,，开始和结束。然后根据这三个信号序列生成建议。最后，PEM预测每个提案的置信度得分，表明提案与最接近的地面真值区间的重叠程度，以决定提案是被接受还是被拒绝。请参阅[23]或我们的补充资料，了解BSN的更多详细信息。



***Figure3：时间扭曲。（a） 通过时间扭曲，我们可以在编码空间中采样更多的片段特征。在这里，我们展示了每个代码段特征的一个简单的二进制分类示例（维度减少到1）。在标记的片段特征（填充圆）中重新采样新的特征点（空圆）可以鼓励学生模型生成更平滑的流形进行预测。（b） 为了进行时间扭曲，我们首先对混合截断正态分布进行采样，以生成一维网格G。然后对特征序列进行网格采样，以增加用于训练的数据。***

### 3.3 平均教师框架(Mean Teacher Framework)

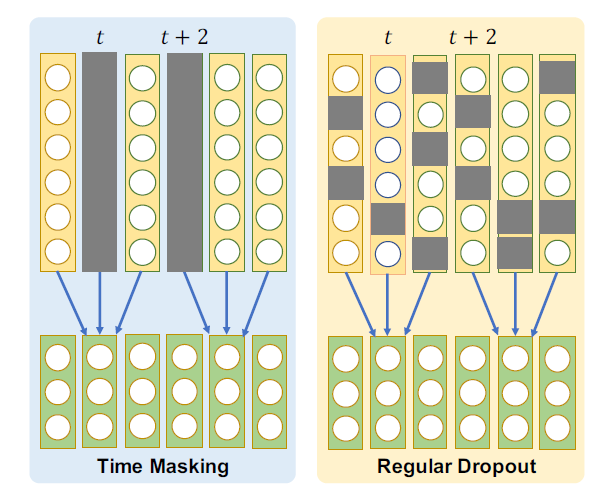
现在，我们将介绍如何为时间动作建议构建半监督学习框架。当只有少量标记的训练样本可用时，像BSN这样的深层模型往往会过度拟合，无法从训练集中提取足够的知识，从而推广到看不见的视频。这可以通过半监督学习来缓解，其中未标记的视频也可以用于培训。如果没有地面真相标签，未标记视频的监督分类损失是不确定的。相反，我们需要引入一个无监督的辅助任务来利用未标记视频中的信息。

作为基线，我们可以直接在时间动作建议模型上调整平均教师方法，形成半监督学习框架。在这个框架中，有两个模型：学生提案模型和教师提案模型。学生在完全监督学习中学习，其权重通过对标记视频应用监督分类损失进行优化。教师建议模型具有与学生相同的神经网络结构，而其权重是通过对不同训练迭代的进行平均来生成的

(1)

其中是平滑系数参数，表示训练迭代。作为集成模型，教师将输入片段特征嵌入到平滑流形中，并输出比学生更一致的预测。然后，无监督的任务是在学生和教师模型的输出之间施加一致性正则化，以标记和未标记的视频作为输入。

### 3.4 连续扰动(Sequential Perturbations)



***Figure4：时间掩蔽。 与将输入中的一些神经元随机归零的 dropout 不同，Time Masking 从随机选择的时间步中删除整个特征向量。***

在Mean Teache框架之外，许多半监督学习工作发现，随机扰动对于学习稳健模型至关重要[21,25,29,36]。扰动的一种典型方式是向特征贴图添加噪声。Mean Teacher[36]将高斯噪声添加到学生和教师模型的中间特征映射中，而VAT[25]将对抗性噪声添加到输入中。在视频分析中，我们进一步探索了顺序学习所需的其他特定干扰。我们提出了两种连续扰动：时间扭曲和时间掩蔽。

**时间扭曲**。 时间扭曲本质上是一个重采样层，它沿着由随机生成的一维流场网格引导的时间维度对一系列特征向量进行重采样。 时间扭曲对于半监督时间动作提议至关重要：首先，通过将标签传播到特征空间中未标记的位置，重采样导致更平滑的预测（图 3（a））； 其次，Time Warping 是一种数据增强的方式，为训练提供更多的标签数据，这在我们标签很少的情况下特别有用； 第三，拉伸和压缩输入信号可以生成更多变体以在某些任务中学习，例如需要准确的开始/结束位置预测的时间动作提议。

为了对输入特征序列执行扭曲，每个输出特征向量是通过根据密集的一维网格 , 应用线性采样来计算的，其中 是对输出特征向量进行采样的时间位置。 执行时间扭曲的关键是，网格应该包括长期失真，它会减慢视频的某些部分，同时加快其他部分的速度； 它还应该包含短期随机噪声。 考虑到这些因素，我们提出了一个混合截断正态分布 (MTND) 采样器（图 3 (b)）来生成网格。

MTND 由 n 个截断正态分布 以不同的权重混合而成。由于我们只想对输入序列进行插值，因此分布在starting (0) 和ending () 位置被截断。 均值 从均匀分布中采样，标准差 从对数均匀分布中采样。 给定一个 MTND，我们从中采样 T 个位置作为网格 G，然后我们执行扭曲并获得 。

**时间掩蔽**。除了时间扭曲之外，我们还提出了一种时间掩蔽操作，作为训练期间顺序扰动的另一个来源。在我们的管道中，时间掩蔽遵循时间扭曲，并将作为输入。时间掩蔽的想法很简单：输入序列中的一些片段被学生模型掩蔽，而教师模型可以看到整个无障碍的视频序列。我们将时间掩蔽的输出表示为。在培训期间，鼓励每次迭代中的蒙面学生模型生成与教师相同的输出，即使他们无法访问输入视频的全部信息。

时间掩蔽可以看作是一个特殊的退出层（图4）。在常规的脱落层，一个片段中的神经元不太可能完全脱落，这使模型有机会从感受野的每个片段中窥视一些信息。相反，在时间掩蔽中，删除的代码段的任何信息都不会传递到下一层。学生模型将被迫从时间上下文中聚合信息，以对删除的片段进行预测。这种时间上下文聚合的能力将从标记视频的监督损失和所有培训数据与教师模型的一致性中学习。

### 3.5 训练(Training)

训练我们的半监督框架包括两部分：最小化标记数据的监督损失和所有训练数据的一致性损失。虽然我们有学生和教师模型，但只有学生模型中的权重通过反向传播优化，而教师模型中的权重是学生的平均权重。

**监督损失**。与完全监督的提案模型一致，我们的半监督框架使用与BSN中相同的监督损失进行培训。有关损失的详细信息，请参阅[23]或我们的补充资料。在我们的半监督框架中，学生提案模型的输出对应于被时间扭曲扭曲的顺序输入。因此，标签也需要根据MTND采样器生成的相同网格重新采样。对于扭曲标签，我们对学生输出施加监督损失。请注意，监督损失只能应用于训练集中的标记视频。

一致性正则化。一致性损失将教师模型的输出视为标签，并鼓励学生学习像教师一样的平滑流形。与监督损失不同，一致性损失可以应用于训练集中有标签和无标签的视频。与我们在监督损失中处理标签的方式类似，我们还将教师的输出扭曲为。然后，一致性损失测量学生输出和扭曲的教师输出之间的距离：

(2)

对于距离函数D，我们在所有实验中使用均方误差。与监督优化一样，只训练学生模型中的权重。一致性损失和监督损失合计为总损失。

## 4 实验

**数据集。** 我们使用 ActivityNet v1.3 和 THUMOS14 进行所有实验。 ActivityNet v1.3 [5] 是一个用于时间动作建议和检测的大型数据库。 它包含 200 个活动类别的 19,994 个视频，并已在 2016 年至 2019 年的 ActivityNet 挑战赛中使用。ActivityNet v1.3 按 2:1:1 的比例分为训练集、验证集和测试集，动作实例的时间边界为 在所有视频中都有注释。 THUMOS14 [18] 分别包含 200 和 213 个时间带注释的未修剪视频，在验证和测试集中分别具有 20 个动作类。 THUMOS14 的训练集是 UCF-101 [35]，其中包含用于动作分类任务的修剪视频。 我们不是在这些修剪过的视频上进行训练，而是在验证集中的未修剪过的视频上训练我们的模型，并在测试集上报告性能。

**评估指标**。 我们在两个任务上评估我们的方法：时间动作建议和时间动作定位。

对于提案，我们报告了每个视频（AN）的各种平均提案数的平均召回率（AR）。 AR 定义为所有召回值的平均值，tIoU 阈值从 0.5 到 1，步长为 0.05。 在 ActivityNet v1.3 上，AR 与 AN 曲线下的面积 (AUC) 也用作度量，其中 AN 从 0 到 100 变化。对于动作定位，我们计算具有不同 tIoU 阈值的平均平均精度 (mAP)。

**实施细节**。 我们遵循与 BSN [23] 相同的预处理和后处理，包括 Soft-NMS 中使用的参数。 对于 ActivityNet v1.3 上的特征提取，我们使用在 Kinetics [20] 上预训练的双流网络 [38]。 与 BSN 的设置不同，我们的特征没有在 ActivityNet 分类任务上进行预训练，以避免使用会污染半监督设置的额外标签。 对于所有 THUMOS14 实验，我们使用与 BSN 相同的视频功能。 对于半监督训练，我们使用 EMA 衰减 α= 0.999。 Time Masking 中的掩蔽概率固定为 0.3。

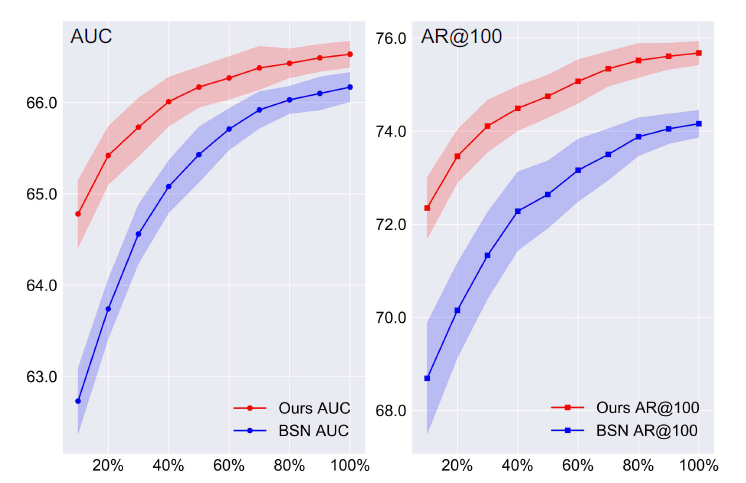
### 4.1 临时行动建议(Temporal Action Proposals)

以一个长的、未修剪的视频作为输入，我们的方法旨在生成时间边界，确定每个动作实例的开始和结束时间。 在本节中，我们将我们的模型在 ActivityNet v1.3 和 THUMOS14 上生成的时间动作提议与全监督 BSN 和其他最先进的方法进行比较，以验证我们的半监督框架的有效性。

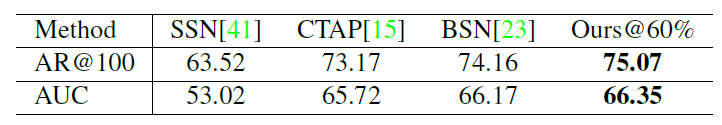
**与全监督方法的比较**。 我们首先在两种训练设置下比较 ActivityNet-1.3 验证集上的行动建议结果：（1）我们的半监督框架，其中 x % 的训练视频标记有时间边界，而 100-x % 的训练视频没有； (2) 最先进的全监督学习，其中使用相同数量的标记视频进行训练，而不使用其他数据。 通过这个比较，我们可以看到我们的半监督框架在不同的标记/未标记比率下与全监督框架的表现如何

为了验证我们方法的标签效率，我们改变了训练标签的数量，然后测量了我们的方法和原始 BSN 生成的提议的 AUC 和 AR@100（图 5）。 由于只有一部分训练集被标记，我们的方法在不同比例的标记训练视频下始终优于完全监督的基线。 值得注意的是，只有 60% 的视频被标记，我们的半监督模型在 AUC 和 AR@100 的两个指标中都优于使用所有标签训练的最先进的全监督 BSN（表 1）。 同样，我们检查 THUMOS14 上的标签效率（图 6），并观察到始终如一的卓越性能。

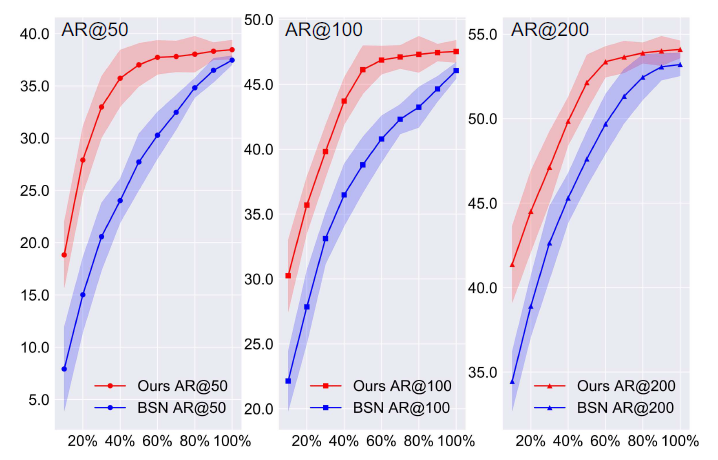
然后，我们将 THUMOS14 上的提案生成与强大的基线模型进行比较。 表 2 显示了在每个视频的不同平均提案数下通过平均召回率测量的比较。 同样，当仅使用 60% 的标签进行训练时，我们的表现优于 BSN。 此外，当 100% 的标签可用时，我们的框架可以进一步提高平均召回率。



***Figure5: 在 ActivityNet v1.3 上进行标签效率实验。 改变训练标签的百分比，我们比较了我们的半监督方法和全监督 BSN 对应方法生成的提议的 AUC 和 AR@100***

******

***Table1: 在 AR@100 和 AUC 方面，我们的方法与 ActivityNet v1.3 上的其他最先进的提案生成方法进行了比较。 在仅使用 60% 的标签时，我们的表现优于所有其他方法。***

******

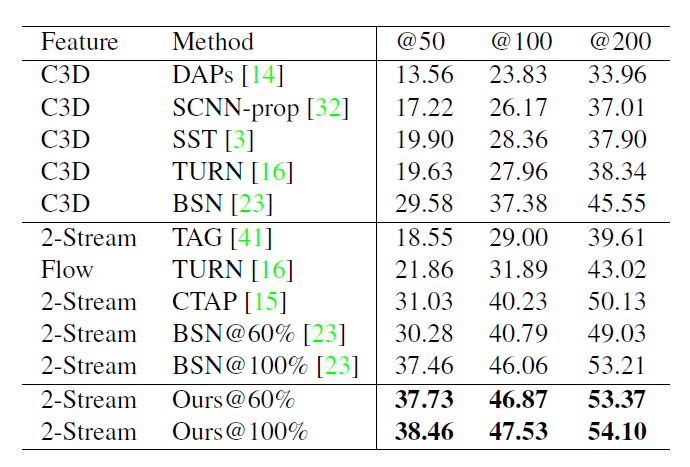
***Figure 6: THUMOS14 上的标签效率实验。 我们报告了 AR@50、@100 和 @200 在使用训练集中不同百分比的标签进行训练时我们的方法和 vanilla BSN 生成的提议***

**与半监督基线的比较**。 接下来，我们针对 THUMOS14 提议的多个半监督基线研究我们的框架的性能，其中 60% 的标签用于训练（表 3）。 我们首先结合 BSN 实施和评估增值税 [25]。 VAT 的关键思想是提高模型对近似最坏情况扰动而不是随机扰动的鲁棒性。 类似于文本分类的增值税应用[24]，我们将对抗性噪声应用于每个视频片段嵌入，而不是直接应用于原始输入。 VAT 并没有显着提高平均召回率，部分原因是视频片段嵌入的最坏情况扰动与随机噪声没有显着差异。

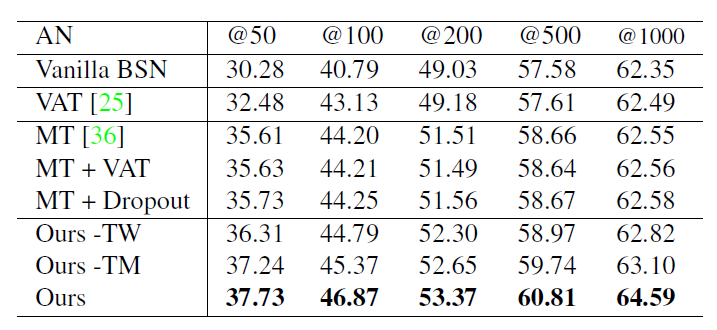
我们还研究了 Mean Teacher [36] 的不同变体。 只有随机噪声且没有 dropout 层的 vanilla Mean Teacher 优于 VAT。 此外，将增值税添加到平均教师对更好的提案没有太大帮助。具有常规 dropout 的 Mean Teacher 进一步提高了提案的质量，但不如我们的 Time Masking 方法强大。 在相同的 dropout/masking 概率下，虽然常规 dropout 在每个训练步骤中将与 Time Masking 相同数量的神经元归零，但它为学生模型制定了更容易学习的任务，因为学生可以依赖更多片段进行推理。

最后，我们通过分别去除它们来检查两个提出的顺序扰动的贡献。他们都对提案做出了贡献，而时间扭曲似乎发挥了重要作用。

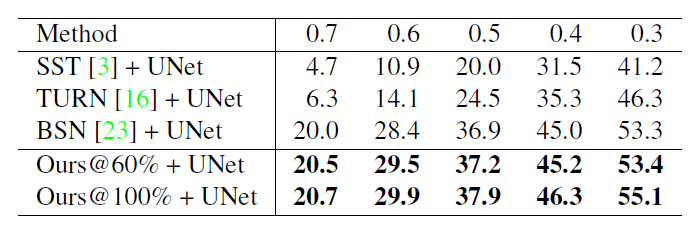
**定性结果**。 我们可视化我们的半监督方法生成的一些时间行动建议。 图 8 表明，当两者都使用 60% 的标签进行训练时，我们的方法能够生成比 THUMOS14 上的完全监督基线更精确的时间边界。



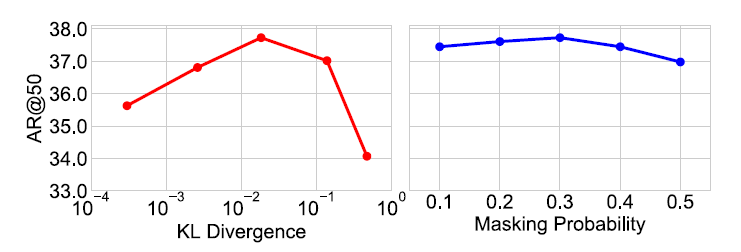
***Table2: 在 AR@50、AR@100 和 AR@200 方面，我们的方法与 THUMOS14 上其他最先进的提案生成方法进行了比较。***

******

***Table3: 使用 60% 的标签训练的全监督和半监督基线之间的比较。 我们在 THUMOS14 上的各种 AN 上报告 AR。 缩写：VAT 表示虚拟对抗训练，MT 表示平均教师，TW 表示时间扭曲，TM 表示时间掩蔽。 我们的完整模型优于强大的半监督基线。***

******

***Table4: 在 THUMOS14 的测试集上以 mAP@tIoU 表示的动作检测结果。 我们与提议+分类方法进行比较，其中分类结果由 UntrimmedNet [33] 生成。***

******

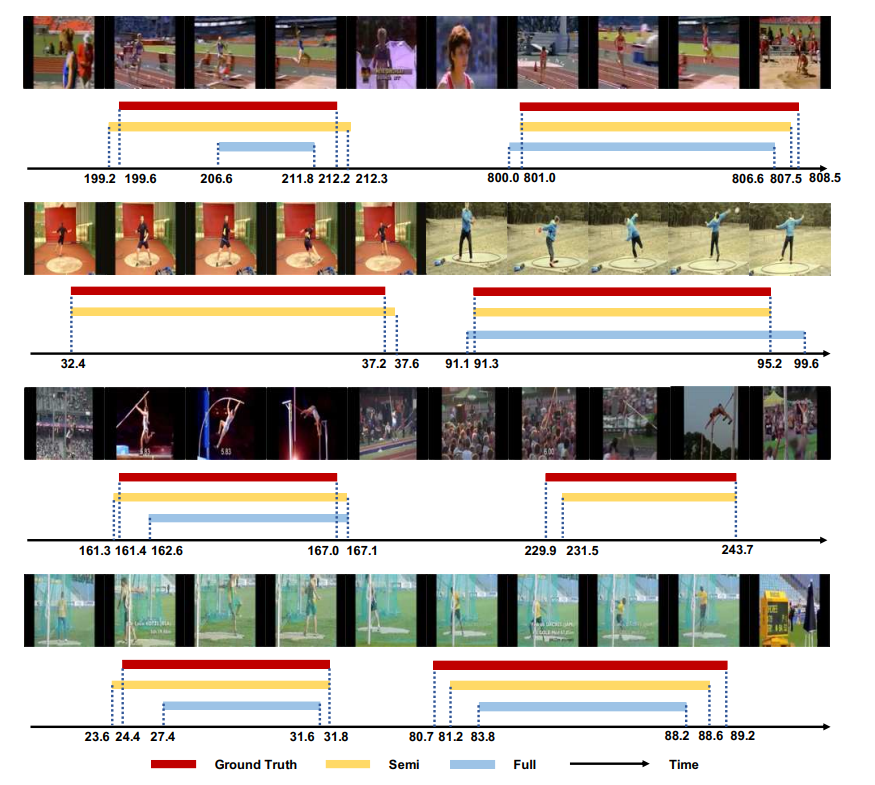
***Figure7: 消融实验。 我们在不同的超参数选择下评估时间扭曲和时间掩蔽的效果，以找到更好的性能的最佳点***

### 4.2 消融实验(Ablation Experiments)

为了评估两个提议的顺序扰动的功能，我们在 THUMOS14 上进行了实验，其中 60% 的标签在时间扭曲和时间掩蔽中使用了不同的超参数.

**TimeWarping 中的失真程度**。 时间扭曲的效果取决于从 MTND 采样器采样的网格。改变截断正态分布的数量及其尺度，MTND 可以从几乎均匀的分布变为非常不均匀的分布，这将极大地扭曲输入序列。 我们研究了时间扭曲中不同程度的失真对生成提案的影响。失真程度由采样 MTND 之间的 KL 散度 测量为 P 和均匀分布 Q。图 7 (a) 显示了 在 0.01 数量级的最佳点。 当 接近 0 时，TimeWarping 的效果减弱； 当失真度太大时，视频的很多部分很难被采样，相当于减少了用于训练的标签数量。

**时间掩蔽中的掩蔽概率**。 我们在输入到 Time Masking 的序列中尝试不同概率的归零特征向量。 如图 7（b）所示，p = 0.3 似乎是一个最佳操作点，给学生带来了适当的难度。 因此，我们在所有实验中都修复了这个掩蔽概率



***Figure8: 我们将我们的半监督方法生成的 THUMOS14 提议与使用 60% 的标签训练的全监督 BSN 进行比较。 我们还显示了真实区间以供参考***

### 4.3 时间动作定位(Temporal Action Localization)

生成时间动作建议的最终目标是时间动作定位，因此我们进一步评估了我们在 THUMOS14 上的定位任务建议。 我们遵循 [3, 16, 23] 中的建议加分类两阶段方法。 正如 BSN 所做的那样，我们使用 UntrimmedNet [33] 在不同方法生成的提案之上预测的前 2 个视频级别类别。 我们在 THUMOS14 上报告了不同时间 IoU 阈值下的平均平均精度，每个视频有 200 个提议（表 4）。 直接比较的是使用所有标签训练的全监督 BSN，我们在从 0.3 到 0.7 的不同时间 IoU 阈值上取得了更好的性能。 当使用所有标签进行训练时，我们的模型进一步提高了动作定位的性能

## 5 结论

我们表明，通过采用我们的半监督方法来学习它们的参数，可以以更高的标签效率训练时间建议模型。 我们的半监督框架扩展了平均教师模型，提出了两个用于视频理解的顺序扰动。我们凭经验证明，当仅使用 60% 的标签进行训练时，我们的模型实现了与全监督方法相似的性能，也优于其他半监督基线。 此外，我们表明我们的半监督建议可以有效地应用于时间动作定位问题。

**致谢**。 这项工作得到了松下和京东美国技术公司（“JD”）在 SAIL-JD AI 研究计划下的支持。 本文仅反映其作者的观点和结论，不代表松下、京东或与松下或京东相关的任何实体。

**引用**

[1] Piotr Bojanowski, R´emi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 3

[2] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei- Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC*, volume 2, page 7, 2017. 2

[3] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6373– 6382. IEEE, 2017. 1, 2, 7, 8

[4] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern* *recognition*, pages 1914–1923, 2016. 2

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern* *Recognition*, pages 961–970, 2015. 2, 5

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE* *Conference on*, pages 4724–4733. IEEE, 2017. 3

[7] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE Conference* *on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019. 3

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. 2

[9] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.;

2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. 2

[10] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. Semi-supervised sequence modeling with cross-view training. In *EMNLP*, 2018. 3

[11] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015. 3

[12] Xuhuan Duan, Le Wang, Changbo Zhai, Nanning Zheng, Qilin Zhang, Zhenxing Niu, and Gang Hua. Joint spatiotemporal action localization in untrimmed videos with perframe segmentation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 918–922. IEEE, 2018. 2

[13] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *Computer Vision, 2009 IEEE 12th International* *Conference on*, pages 1491–1498. IEEE, 2009. 2

[14] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer* *Vision*, pages 768–784. Springer, 2016. 1, 2, 7

[15] Jiyang Gao\*, Kan Chen\*, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *ECCV*, 2018. 1, 2, 6, 7

[16] Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals, 2017. 2, 7, 8

[17] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer, 2016. 3

[18] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. 5

[19] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. Fast saliency based pooling of fisher encoded

dense trajectories. In *ECCV THUMOS Workshop*, volume 1, page 7, 2014. 2

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 6

[21] Samuli Laine and Timo Aila. Temporal ensembling for semisupervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1, 2, 4

[22] 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. 2

[23] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on* *Computer Vision*, 2018. 1, 2, 3, 4, 5, 6, 7, 8

[24] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarialtraining methods for semi-supervised text classification. In *ICLR*, 2017. 3, 7

[25] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE* *transactions on pattern analysis and machine intelligence*, 2018. 1, 2, 3, 4, 7

[26] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *Proceedings of the IEEE Conference on Computer* *Vision and Pattern Recognition*, pages 1020–1028, 2016. 2

[27] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. The lear submission at thumos 2014. 2014. 2

[28] Isabeau Pr´emont-Schwarz, Alexander Ilin, Tele Hao, AnttiRasmus, Rinu Boney, and Harri Valpola. Recurrent ladder networks. In *Advances in Neural Information Processing* *Systems*, pages 6009–6019, 2017. 3

[29] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015. 1, 2, 4

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2

[31] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weaklysupervised temporal action localization in untrimmed videos. In *ECCV*, pages 162–179, 2018. 3

[32] Zheng Shou, DongangWang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision* *and Pattern Recognition*, pages 1049–1058, 2016. 2, 7

[33] Gurkirt Singh and Fabio Cuzzolin. Untrimmed video classification for activity detection: submission to activitynet challenge. *arXiv preprint arXiv:1607.01979*, 2016. 2, 7, 8

[34] Khurram Soomro and Mubarak Shah. Unsupervised action discovery and localization in videos. In *Proceedings of the* *IEEE Conference on Computer Vision and Pattern Recognition*, pages 696–705, 2017. 1

[35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5

[36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural* *information processing systems*, pages 1195–1204, 2017. 1, 2, 3, 4, 7

[37] LiminWang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2):2, 2014. 2

[38] 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 *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 3, 6

[39] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kassim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference* *on Computer Vision and Pattern Recognition*, pages 3093– 3102, 2016. 2

[40] Ze-Huan Yuan, Jonathan C Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *CVPR*, volume 2, page 7, 2017. 2

[41] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. *ICCV, Oct*, 2, 2017. 2, 6, 7

[42] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.2