

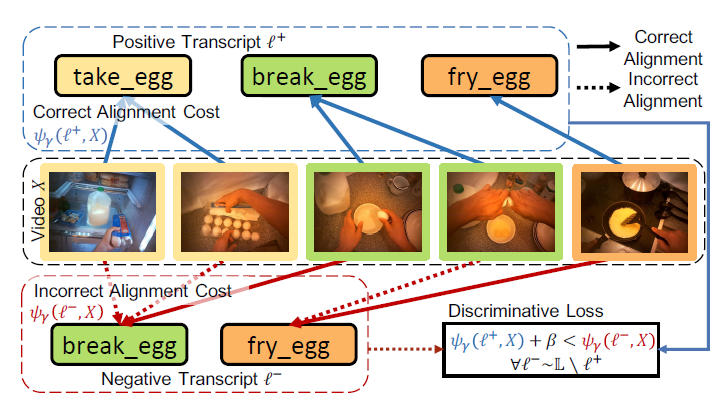


**DTW：用于弱监督动作对齐和分割的判别可微动态时间规整**

## 摘要

我们解决了视频中弱监督的动作对齐和分割问题，其中在训练期间只有发生动作的顺序可用。 我们提出了判别可微动态时间规整 (DTW)，这是第一个使用弱排序监督的判别模型。 具有弱监督的判别建模的关键技术挑战是排序监督的损失函数通常使用动态规划制定，因此不可微。 我们通过在动态规划中不断放宽最小算子来解决这一挑战，并将对齐损失扩展为可微分。 所提出的 DTW 通过判别建模和端到端训练创新地解决了序列对齐问题，从而显着提高了弱监督动作对齐和分割任务的性能。 我们表明，我们的模型能够绕过以前工作中通常遇到的退化序列问题，并在两个具有挑战性的数据集中的三个评估指标上优于当前最先进的模型。

## 1 介绍



***Figure1： 我们只使用有序的动作列表或成绩单作为训练的弱监督。 此设置具有挑战性，因为所需的输出在训练时不可用。 我们通过为这项任务提出第一个判别模型来应对这一挑战。 将视频 （中间）对齐到地面实况或正转录本 （顶部）的成本应该小于随机采样的负转录本（底部）的成本。***

近年来，由于视频数据量大，视频动作理解越来越受到关注。 与需要注释每个动作的确切开始和结束时间的完全注释方法 [20, 31, 39] 相比，弱监督方法 [7, 15, 29, 3, 21] 显着减少了所需的注释工作并改进了 对真实世界数据的适用性。 特别是，我们关注一种通常称为动作顺序或转录本的弱标签，它使用视频中发生的动作的有序列表作为监督。

仅使用动作顺序作为监督的主要挑战是在训练时不提供地面实况目标、逐帧动作标签。 以前的工作求助于使用各种替代损失函数，这些函数可以最大化弱标签的后验概率或给定视频的动作排序。 然而，如 [15] 所示，使用代理损失函数很容易导致退化的结果，将一些发生的动作与视频中的单个帧对齐。 这种退化的结果与我们期望的基本事实相去甚远，因为每个动作在执行过程中通常跨越许多帧。 虽然以前的工作试图使用帧间相似性 [15]、细到粗策略 [28] 和段长度建模 [29] 来解决这一挑战，但这些方法仍然考虑与单帧对齐的退化结果 作为受代理损失函数约束的有效解决方案。

本文的主要贡献是通过提出第一个使用订单监督的判别模型来应对挑战。 如图 1 所示，这个想法是与正面或基本事实转录本正确对齐的概率应该高于负面转录本的概率。 与之前仅最大化弱标签的后验概率的工作相比[15,28,29]，我们的判别公式不会受到退化对齐的影响，因为它不再是新提出的判别损失的明显且微不足道的解决方案。 此外，与以前的工作相比，最小化判别损失直接有助于改进我们的目标。 其他研究领域也研究了类似的想法，例如用于图像标记的多实例学习，并已被证明是成功的 [37]。

虽然将判别建模应用于弱监督动作标记问题的想法看似直观，但关键的技术挑战是，先前方法中损失函数的计算通常涉及不可微的结构预测算法，例如动态规划 (DP)。 我们通过提出 Discriminative Differentiable Dynamic TimeWarping (D3TW) 来解决这一挑战，我们通过最小化通过连续松弛 DP [26] 中的最小算子获得的判别损失函数来直接优化以获得更好的输出。 D3TW 的使用使我们能够将判别建模的优势与结构预测模型结合起来，这在以前的方法中是不可能的。

我们在早餐行动 [20] 和好莱坞扩展 [3] 两个流行的基准数据集中的两个弱监督任务上评估 D3TW。 第一个任务是动作分割，它指的是预测逐帧动作标签，其中测试视频是在没有任何进一步注释的情况下给出的。 第二个任务是动作对齐，正如 [3] 中提出的，它是指将测试视频序列与给定的动作顺序序列对齐。 我们展示了我们的 D3TW 显着提高了这两个任务的性能。

总之，我们的主要贡献是：（i）我们引入了第一个用于排序监督的判别模型来解决退化序列问题。 (ii) 我们提出了 D3TW，这是一种新颖的框架，它结合了判别建模和端到端训练的优势，用于具有弱监督的结构序列预测。 (iii) 我们将我们的方法应用于两个具有挑战性的现实世界视频数据集，并表明它在弱监督动作分割和对齐方面都达到了最先进的水平。

## 2 相关工作

**动作识别和分割**。 动作识别一直是视频理解的一项重要任务[13,27,33,36]。 随着修剪视频数据集的性能提高 [13, 4]，最近视频理解的重点已经转移到更长和未修剪的视频数据上，例如 VLOG [10]、Charades [35] 和 EPIC-Kitchens [6]。这导致了动作分割方法 [23, 34, 39] 的发展，旨在标记视频中的每一帧，而不仅仅是对修剪的视频剪辑进行分类。 我们的目标也是密集标记视频的每一帧，但没有密集的训练监督。

**视觉中的弱监督学习**。 对于图像，弱监督学习已在分类 [37, 24]、语义分割 [40]、对象检测 [22] 和视觉基础 [17, 38] 中进行了研究。 排序约束已被广泛用作视频中的弱监督 [3, 2, 7, 15, 28, 29]。 与我们的工作最接近的是 NN-Viterbi [29]，它结合了一个神经网络和一个不可微的 Viterbi 过程，以迭代地从排序监督中学习。 相比之下，所提出的 D3TW 是端到端可微的，并使用判别建模直接优化排序监督下的最佳对齐。

**使用语言作为视频的监督**。 由于排序监督可以从语言中自动提取，我们的工作涉及使用语言作为视频的监督。 监督通常来自电影剧本 [8, 2, 41] 或教学视频的转录 [1, 33, 25, 14]。 与这些方法不同，我们假设已经提取了离散的动作标签，并专注于利用排序信息作为监督。

**持续放松**。 我们的 D3TW 与离散操作的连续松弛的最新进展有关，包括定理证明 [30]、softmax 函数 [16]、逻辑规划 [9] 和动态规划 [26, 5]。 我们使用相同的原理并进一步启用基于动态规划的对齐的判别建模。

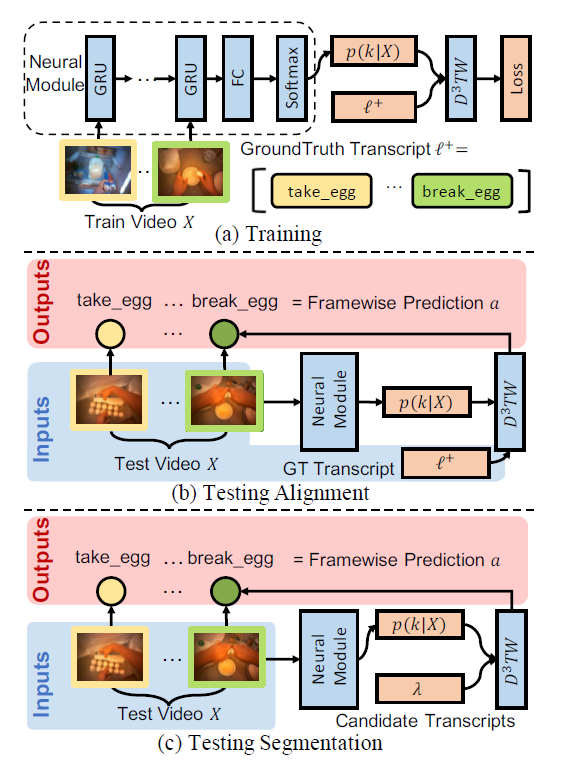
## 3 方法

我们的目标是学习仅使用弱监督在时间上对齐和分割视频帧，其中只有发生动作的顺序在训练时可用。 弱监督问题的主要挑战是地面实况目标，即逐帧动作标签在训练时不可用。 我们通过提出判别可微动态时间规整 (D3TW) 来应对这一挑战，据我们所知，这是第一个具有排序监督的判别建模框架。判别建模和可微动态规划的使用使我们的方法与以前涉及不可微前向后算法 [11,15,29] 的工作不同，并显着缓解了将每个动作标签对齐到单个帧的退化对齐问题。 图 2 显示了我们模型的轮廓。下面，我们从问题陈述开始详细描述我们的框架。 然后我们定义我们的模型并展示如何在测试时使用它。

### 3.1 弱监督行动学习

我们从弱监督动作对齐和分割的定义开始。 这里的弱监督意味着在训练时只提供成绩单或动作的有序列表。 例如，煎鸡蛋的视频可能包括取鸡蛋、打鸡蛋和煎鸡蛋。 虽然完全监督将提供每个动作的细粒度时间边界，但在我们的弱监督设置中，只给出了动作顺序序列 [take\_egg, break\_egg, frog\_egg]。

我们在本文中解决了两个任务：动作分割和动作对齐。 我们的目标是在弱监督下学习两者。 如图 2(b) 和 (c) 所示，这两个任务之间的区别在于，在测试时，动作对齐使用脚本和测试视频帧作为输入，而动作分割只需要测试视频帧作为输入。我们观察到，在测试时给定一组可能的转录本，动作分割可以表述为动作对齐任务。 我们将首先解释如何使用弱监督来解决动作对齐问题，并解释如何将动作分割简化为动作对齐问题。



***Figure2 (a) 在训练期间，只给出成绩单 。 输入视频首先通过 GRU 转发，以生成每帧每个动作的后验概率 。 D3TW 是一个具有完全可微损失函数的判别模型，它允许我们通过反向传播学习 ，并将我们的方法与以前的工作区分开来。 (b) 对于对齐，在测试时，我们的 D3TW 损失可以直接用于将给定的成绩单 与视频序列对齐。 (c) 对于分段，在测试时没有给出成绩单。 我们通过将视频与一组候选转录本对齐来减少分割到对齐，并输出最佳候选作为分割结果。***

形式上，给定视频帧的输入序列 ，动作对齐的目标是预测逐帧动作标签的输出对齐序列 ，在 遵循转录中的动作顺序的约束条件下 。 这里， 是一组可能的动作。 换句话说，我们想学习一个模型 。 弱监督的关键挑战是我们只有输入 作为训练 的监督，而无法访问地面实况动作标签 。

对于动作分割，我们观察到分割可以表述为给定一组可能的转录本的对齐。 形式上，给定一组可能的转录本，让 是一个评分函数，用于衡量给定输入视频 的预测动作标签的好坏，动作分割任务可以通过穷举搜索来解决

(1)

这会找到候选转录本 ，它给出了 中转录本的最佳比对，由 测量。

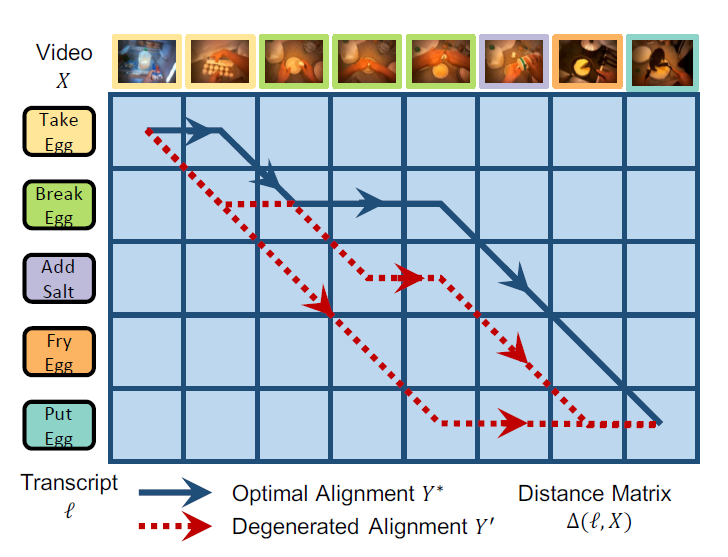
### 3.2 判别可微 DTW (D3TW)

我们已经讨论了什么是弱监督动作对齐以及我们如何解决基于对齐的动作分割。 现在我们讨论如何使用判别建模来学习一个模型，该模型在训练时仅使用 和 来对齐转录 和视频帧 。

我们将动作对齐作为动态时间规整 (DTW) [32] 问题提出，该问题已广泛应用于语音识别中的序列对齐。 给定一个距离函数 来衡量将帧 与转录本 中的标签对齐的成本，DTW 使用动态规划来有效地找到使总成本最小化的最佳对齐方式。 弱监督学习的关键挑战是没有帧到帧的对齐标签来训练这个距离函数 。 我们通过提出判别可微动态时间规整 () 来应对这一挑战，它允许我们仅使用弱监督来学习 。 在下文中，我们将首先讨论如何将视频对齐公式化为 DTW，然后我们将如何使用 学习距离函数

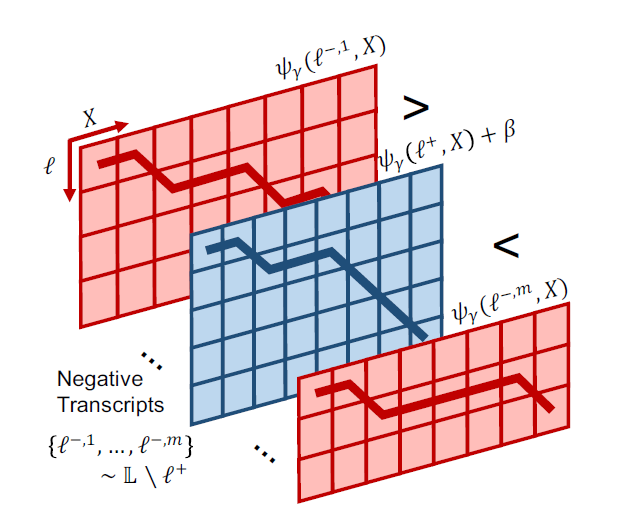
#### 3.2.1 视频对齐作为动态时间扭曲

给定两个序列 和 的长度 和 对应于转录和视频，我们将 定义为可能的二进制对齐矩阵的集合。 这里 ，如果视频帧 被标记为，则 = 1，否则 = 0。 基于观察到每个视频帧只能与单个动作标签对齐，我们对合格的扭曲路径施加严格的约束，使得从 到 的对齐严格是一对一的。换句话说，是具有恰好 T 个非零元素和列主元的二进制矩阵的集合。 给定一个对齐矩阵 Y，我们可以将其对应的动作标签 推导出为： ，如果 。



***Figure3： 用于视频对齐的动态时间扭曲公式。 彩色网格表示距离矩阵。***

***在这里，我们使用格子图来显示最佳转录视频对齐 的计算图，如Eq(2)所定义。贝尔曼递归保证 和成绩单中的动作顺序被严格保留。***

******

***Figure4： 我们将判别建模引入弱监督动作对齐。 将视频 与正确转录本 对齐的损失***   ***应该低于任何其他随机样本负转录本 的损失，这可以防止以前工作中常见的退化对齐问题。***

考虑到对合格路线的限制，DTW的目标是找到最佳路线

这使得对齐矩阵和在转录本和视频之间的距离矩阵之间的内积最小化，其中

给定距离函数，我们通过动态规划可以解决Eq（2）。图3显示了这种过程的简化示例。在所有将左上角入口连接到右下角入口的路径中，仅使用移动，是使转录序列和视频帧之间的对齐成本最小化的最佳对齐。在这种情况下，我们可以有效地获得视频X和转录本之间的最佳对齐。

#### 3.2.2 弱监督下的判别建模

我们已经讨论了在给定使用DTW的距离函数如何获得最佳对齐。然而，问题仍然是，我们如何才能在不接触地面真相对齐的情况下学习这种距离函数。在之前的工作[15,28,29]中使用的一种方法最大限度地提高了视频X在给定成绩单的情况下出现的概率：

(3)

其中 是第帧的动作标签。 通过Eq. (3)中的目标。 我们可以学习，即在给定动作的情况下观察 的概率。为了最大化概率，我们定义距离 作为负对数似然。

人们应该注意到Eq. (3)中的对齐是潜在的，可能的对齐数量随着视频的长度呈指数增长。 因此，以前的工作要么使用动态规划 [15]，要么使用硬 EM (hard EM)方法 [28, 29] 来推断并迭代最大化等式中的目标。 (3)。 这种方法的主要缺点是，由于对齐空间太大，它们很容易导致退化或琐碎的解决方案。 虽然可以通过对可能的对齐 强制执行启发式先验来施加约束，但这并不能直接解决最大化该目标不一定会导致正确对齐的缺点。

我们在这里的主要见解是将判别建模引入弱排序监督问题。 我们强制执行一个判别约束，该约束应该适用于任何输入元组 ，即

(4)

如图 4 所示，根据基本事实或正面转录本 观察视频的概率应始终高于从负面转录本观察视频的概率 . 使用边距为的铰链损失，损失函数可以写成：

(5)

#### 3.2.3 连续松弛的可微损失

虽然上述判别建模很直观，但Eq(5)中的技术挑战是 和 。 对于我们要学习的距离函数 通常不可微。 优化它的一种方法是使用硬 EM 并在给定当前距离函数 的情况下迭代优化此损失。 然而，硬(hard) EM 在数值上是不稳定的，因为它在交互中使用硬最大算子来更新模型参数 [26]。 我们方法的关键技术贡献是提出了对基于 DTW 的视频对齐损失函数的持续放宽。

而不是通过求解Eq(2)来迭代更新模型参数。 给定当前 与硬(hard) EM 找到最佳对齐，我们可以解决以下连续松弛(continuous relaxation)：

(6)

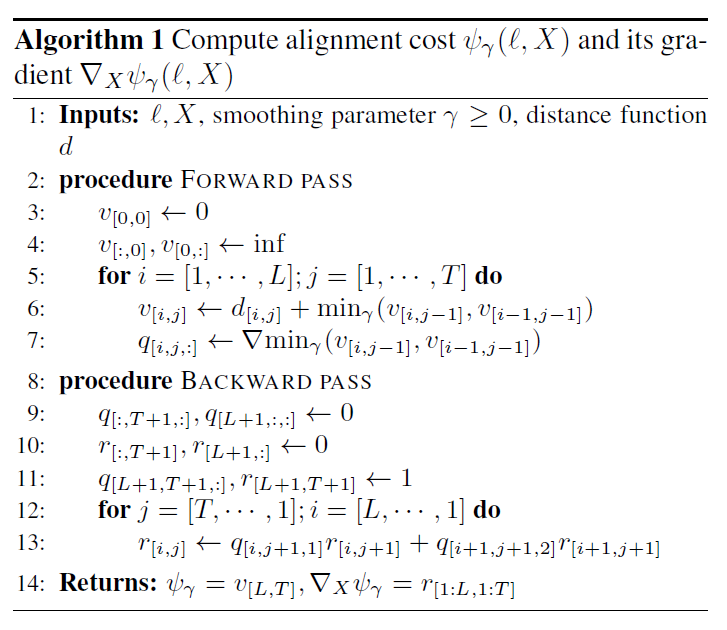
这里 是由负熵 正则化的正则最小算子的连续松弛，平滑参数为 ，使得

(7)

当 时，这会将基于动态规划的 DTW 损失函数转换为关于 的可微分函数。平滑参数 在经验上有助于优化，尽管它没有显式地凸化目标函数。 Eq(6)的梯度可以使用链式法则推导出：

(8)

其中右边第二项可以解释为吉布斯分布 下的平均对齐矩阵。算法 1 总结了计算 及其梯度的过程。



我们可以翻译作为转录本和视频之间所有可能对齐的预期成本。其梯度可以看作是硬对齐的放松版本 在式（2）中。利用式（6）中的连续松弛，我们可以直接计算梯度，并对式（5）进行优化。这解决了由于硬EM中的数值不稳定操作而导致退化对齐的挑战。用我们的松弛对齐成本替换等式（5）中的，我们得到了判别和可微损失函数：

(9)

直接最小化等式（9）使我们的模型能够同时优化，以便在给定观察到的视频序列的情况下找到最佳对齐方式并识别最准确的转录本。等式（9）的可微性允许梯度在整个模型中反向传播，并通过端到端训练微调校准任务中距离矩阵的距离函数。

#### 3.2.4 学习与推断(Learning and Inference)

**距离函数参数化**。在本文中，我们使用一个带有softmax输出层的递归神经网络（RNN）来参数化给定视频帧作为输入的距离函数。设是每帧的RNN输出，其中是可能的操作数。可以解释为时间t时作用k的后验概率。我们遵循[29]并近似计算发射概率，其中是作用类的前验概率。动作类优先级统一初始化为，并在每一批迭代后通过计算和规范化训练过程中迄今为止已处理的每个动作类的出现次数来更新。

**动作分割推理**。在测试时，我们希望我们的模型预测出最佳的动作标签，只给出一个看不见的测试视频。

我们将动作分割任务分解为两部分：首先，我们生成一组候选转录本 [29]后面的表示所有可能的转录本。然后，我们将每个候选成绩单与看不见的测试视频对齐，以找到将对齐成本降至最低的成绩单

预测的对齐和相关的帧级动作标签由。

## 4 实验

的主要贡献是在弱标签和视频帧之间应用区分性、可微性和动态对齐。在本节中，我们将在两个具有挑战性的弱监督任务上评估所提出的模型：两个真实数据集中的动作分割和对齐。此外，我们还研究了我们的模型的分割性能如何随着更多的监督而变化。通过消融研究，我们进一步研究了的有效性，并将我们的方法与目前最先进的方法进行了比较。

**数据集和功能**。**早餐行动**[20]包括1712个未经剪辑的视频，52名参与者在18个不同的厨房烹饪10道菜，如煎蛋。总的来说，大约有360万个框架标记了48种可能的动作。该数据集已被广泛用于弱监督行为标记[7,15,28,29]。为了进行公平比较，我们使用[20]提供的预

计算特征和数据分割。《**好莱坞扩展版**》[3]由937个视频组成，每个视频中包含2到11个动作。总的来说，大约有0.8米的框架标有16种可能的动作，比如开门。我们使用该功能并遵循[3]中的数据分割，以进行公平比较。

**网络架构**。我们使用带有512个隐藏单元的单层GRU[12]。我们使用Adam[18]进行优化，并交叉验证学习率和批量大小等超参数。

**帧子采样**。为了更快地进行训练和推理，我们在早餐动作中临时对特征向量进行子采样。

在[15]之后，我们使用k-均值对视觉上相似且时间上相邻的帧进行聚类，其中中心作为初始化在时间上均匀分布。根据经验，我们选择M=20，这比平均作用时间短得多(⇠早餐数据集中有400帧）。好莱坞扩展数据集不需要进一步的预处理，因为特征向量已经过亚采样。

基线。我们比较了以下六条基线：-ECTC[15]不依赖硬EM。但是，它使用基于不可微DP的算法来计算其梯度。此外，它还包括类之间上下文的显式模型。

-*GRU reest.*。[28]使用隐马尔可夫模型，迭代训练他们的系统，以重新估计输出。

-*TCFPN* [7]也基于动作对齐。然而，它使用了一个迭代框架，既不像那样是可微的，也不像那样是有区别的。

-*NN-Viterbi* [29]与我们的最为相似，可以被视为一种无差别建模且无可微分损失的烧蚀。然而，我们的RNN将整个视频作为输入，而不是视频片段。

-*Ours w/o D*3*TW*是我们的模型，不使用，而是使用类似于NN Viterbi[29]的迭代策略。

这显示了我们的模型在没有区分和可微建模的情况下的性能。

-*Ours w/o Discriminative*模型被比较，以显示判别模型在弱监督学习中的重要性。与我们的不含的模型相比，该模型以等式（3）的可微松弛为目标。



***Figure6： 定性结果表明了判别建模的重要性。我们计算 Facc.，我们的和我们的w/o之间帧精度的绝对差异。有区别的建模能够改善早餐数据集中几乎所有食谱或活动的性能。煎饼（第3行）和炒蛋（第4行）在没有取得显著改善的情况下，我们从更进一步的角度看到了烹饪步骤极其相似的挑战。当烹饪步骤不同时，如三明治（第1行）和谷物（第2行），我们的能够大幅提高帧精度20%以上。***

### 4.1 弱监督动作分割

在分割任务中，目标是预测看不见的测试视频的帧级动作标签，而不进行任何注释。弱监督的动作分割具有挑战性，因为目标输出从未用于训练。如第3.2.4节所述，我们将分割任务简化为对齐任务，首先在给定一组候选转录本的情况下，找到预测转录本使等式（10）中的可能性最大化，然后从和视频之间的对齐中得出逐帧标签。为了公平比较，我们遵循[29]，将设置为训练时间内看到的所有成绩单的集合。

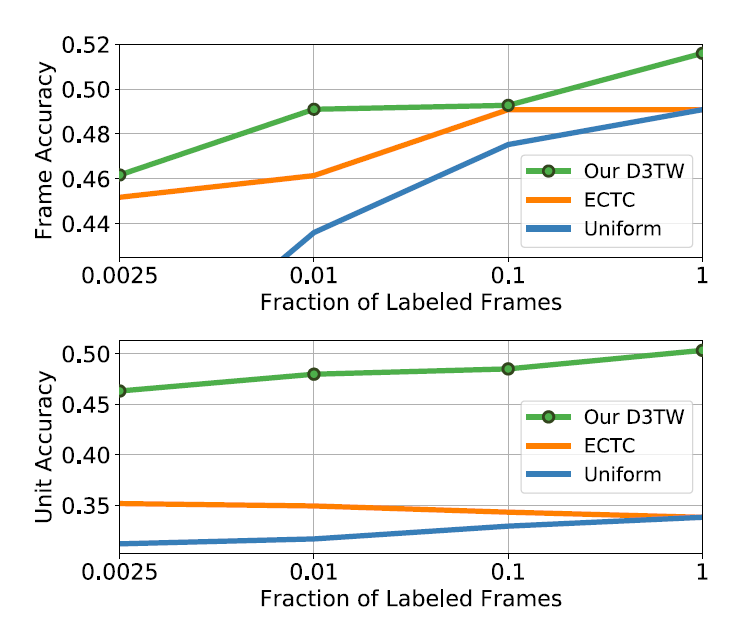
**指标**。我们遵循之前工作[20]中使用的指标来评估预测的帧式动作标签。

第一个是帧精度，即正确标记的帧的百分比。第二个是单位精度，这与语音识别中的单词错误率相似[19]。在计算错误率之前，首先通过动态时间扭曲（DTW）将输出动作标签序列与地面真值标签序列对齐。

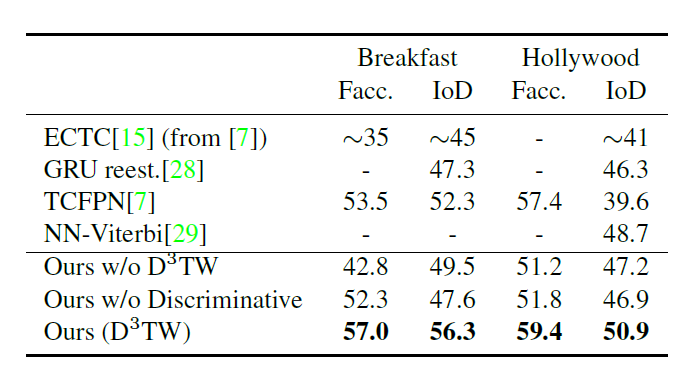
**结果**。弱监督动作分割的结果如表1所示。首先，通过对类之间的上下文及其时间进程进行显式建模，GRU-reest[28]和NN-Viterbi[29]都能够在很大程度上优于ECTC[15]。此外，我们可以看到，基于TCFPN[7]，使用对齐是一种有效的策略。与最先进的方法相比，我们的能够结合这些优势，并表现得相当好。我们的w/o判别通过使用数值稳定性更好的可微松弛损失函数，进一步改进了所有度量。最重要的是，我们使用的完整模型能够将可微损失的好处与区分性建模结合起来，显著优于所有基线，并在所有指标上实现最先进的结果。这表明了我们提出的模型的两个组成部分的重要性。图5显示了视频制作三明治上模型的定性比较。颜色表示不同的动作，横轴是时间。我们的是唯一一个通过区分建模正确捕获所有发生动作的模型。此外，这也会导致更准确的行动边界。通过比较NN Viterbi和我们的w/o判别模型，可以看出可微模型的优点，从而获得更好的动作边界。此外，我们在图6中通过比较我们的完整模型和我们的w/o Discritive模型进一步说明了判别建模的重要性，并显示了我们模型的正确预测、误报和误报。如图所示，区分性建模几乎改善了早餐数据集中的所有10道菜，只有炒蛋例外，其的帧精度较低，可忽略0.2%。我们可以看到，对于我们的没有太大改善的煎饼和炒蛋的菜肴或活动，假阳性在视觉上与正确的预测非常相似，并导致将视频与转录本对齐的挑战。另一方面，对于三明治和谷物等涉及不同步骤的活动，我们的显著地将模型的性能提高了20%以上的帧精度。此外，如果我们观察谷物的假阳性，它只会失败，因为从一个受阻的视角，从视觉上很难区分倒入谷物和倒入面粉的类似动作。

### 4.2 半监督动作分割

与大多数基线相比，我们基于DTW的弱监督动作对齐公式可以通过在计算动作时施加路径约束，轻松地纳入任何额外的帧监督！。这也称为帧级半监督设置，如[15]所述。在半监督设置中，视频中只有少数帧被稀疏地用地面真实感动作注释，这对于注释者来说更容易注释。在这种情况下，我们只与ECTC进行比较，因为它是唯一允许进行该实验的基线。我们进一步比较了[15]中讨论的“统一”基线，其中模型使用了通过按照顺序均匀分布转录本生成的伪标签。帧级半监督动作分割的结果如图7所示。我们可以看到，所提出的在半监督环境下也能够显著提高性能。这再次表明了可微损失函数和判别建模的重要性。



***Figure7： 在早餐数据集的帧级半监督设置中，帧和单位精度与一小部分标记数据相对应。我们基于DTW的公式允许在动态规划中容易地将帧级监控作为路径约束合并。即使在半监督环境下，我们的可微和区分建模也能够在这两个指标上获得更好的性能。***

******

***Table2： 弱监督的动作对齐结果。与分割相比，对齐时给出了地面真实记录，因此性能更高。尽管如此，可微松弛和判别建模仍然有利于这项任务，并导致最先进的结果。***

### 4.3 弱监督动作对齐

在这个任务中，目标是将给定的转录本与其在测试视频中的适当时间位置对齐。我们的配方旨在直接优化行动一致性，只需微弱的监督。在这种情况下，我们总是有基本事实记录，不必使用公式（10）进行搜索。值得注意的是，对齐的结果可以解释为我们的模型在动作分割中表现的经验上限。

韵律学。本实验的主要目标是评估我们的模型，使地面真相记录与输入视频帧对齐。我们使用帧精度等指标来测量预测中的精确时间边界。我们降低了单位精度，因为它使用DTW不可避免地混淆了确切的时间边界。除了帧精度外，我们还使用交叉点过检测（IoD）来测量对齐质量，如下[3]。给我一个真实的行动间隔 预测区间，IoD被定义为。读者应该注意，IoD有时被称为Jaccard度量[3,29]。IoD的值在0到1之间，越高越好。我们报告了测试集中所有地面真值区间的平均IoD。

**结果：**弱监督动作对齐的结果如表2所示。我们可以看到，所有基线的性能在帧精度方面都有所提高，这是因为我们在测试时有更多关于视频在动对齐的信息。这也意味着不同方法之间的差距可能更小。然而，我们观察到的趋势与在行动细分中看到的趋势相同，即拟议的能够在指标上显著优于所有基线，并实现最先进的结果。这个实验再次验证了使用可微损失和区分性建模对我们模型的成功很重要。

## 5 结论

我们提出了，这是第一个用于弱监督动作对齐和分割的鉴别框架。我们工作的关键观察是在阳性和阴性转录本之间使用区分性建模，并绕过退化序列的问题。主要的挑战是基于动态规划的损失通常是不可微的。我们提出了一种持续的放松方法，允许通过端到端训练直接优化判别目标。我们的结果和消融研究表明，区分性建模和可微松弛对的成功至关重要，D3TW在两个具有挑战性的真实数据集上实现了最先进的分割和对齐结果。我们的D3TW框架是通用的，可以扩展到其他需要输出中的优先结构和端到端可微性的任务。

**致谢**。这项工作部分由丰田研究院（TRI）资助。本文仅反映作者的观点和结论，而非TRI或任何其他丰田实体。

**摘要**

[1] J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev, and S. Lacoste-Julien. Unsupervised learning from narrated instruction videos. *CVPR*, 2016. 2

[2] P. Bojanowski, R. Lagugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *ICCV*, 2015. 2

[3] P. Bojanowski, R. Lajugie, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Weakly supervised action label-

ing in videos under ordering constraints. In *ECCV*, 2014. 1, 2, 6, 8

[4] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[5] M. Cuturi and M. Blondel. Soft-dtw: a differentiable loss function for time-series. In *International Conference on Machine Learning*, pages 894–903, 2017. 2

[6] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett,

W. Price, and M. Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[7] L. Ding and C. Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern* *Recognition*, pages 6508–6516, 2018. 1, 2, 6, 7, 8

[8] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009.2

[9] R. Evans and E. Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018. 2

[10] D. F. Fouhey, W.-c. Kuo, A. A. Efros, and J. Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 2

[11] A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 2

[12] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5):602–610, 2005. 6

[13] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[14] D.-A. Huang\*, S. Buch\*, L. Dery, A. Garg, L. Fei-Fei, and J. C. Niebles. Finding “it”: Weakly-supervised, reference-aware visual grounding in instructional videos. In *IEEE* *Conference on Computer Vision and Pattern Recognition* *(CVPR)*, 2018. 2

[15] D.-A. Huang, L. Fei-Fei, and J. C. Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer, 2016. 1, 2, 4, 6, 7, 8

[16] E. Jang, S. Gu, and B. Poole. Categorical reparametrization with gumble-softmax. In *ICLR*, 2017. 2

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.2

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[19] D. Klakow and J. Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002. 7

[20] H. Kuehne, A. Arslan, and T. Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, 2014. 1, 2, 6, 7

[21] H. Kuehne, A. Richard, and J. Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and* *Image Understanding*, 163:78–89, 2017. 1

[22] K. Kumar Singh, F. Xiao, and Y. Jae Lee. Track and transfer: Watching videos to simulate strong human supervision for weakly-supervised object detection. In *CVPR*, 2016. 2

[23] C. Lea, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In *European Conference on Computer Vision*, 2016.2

[24] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. *arXiv preprintarXiv:1805.00932*, 2018. 2

[25] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabi-novich, and K. Murphy. What’s cookin’? interpreting cook-ing videos using text, speech and vision. *NAACL*, 2015. 2

[26] A. Mensch and M. Blondel. Differentiable dynamic programming for structured prediction and attention. *ICML*,2018. 2, 5

[27] H. Pirsiavash and D. Ramanan. Parsing videos of actions with segmental grammars. In *CVPR*, 2014. 2

[28] A. Richard, H. Kuehne, and J. Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *IEEE Conf. on Computer Vision and Pattern Recognition*,volume 1, page 3, 2017. 1, 2, 4, 5, 6, 7, 8

[29] A. Richard, H. Kuehne, A. Iqbal, and J. Gall. Neuralnetwork-viterbi: A framework for weakly super-

vised video learning. In *IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, 2018. 1, 2, 4, 5, 6, 7, 8

[30] T. Rockt¨aschel and S. Riedel. End-to-end differentiable proving. In *NIPS*, 2017. 2

[31] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, 2012. 1

[32] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech* *and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978. 3

[33] O. Sener, A. Zamir, S. Savarese, and A. Saxena. Unsuper- vised semantic parsing of video collections. In *ICCV*, 2015.2

[34] G. A. Sigurdsson, S. Divvala, A. Farhadi, and A. Gupta. Asynchronous temporal fields for action recognition. In *The* *IEEE Conference on Computer Vision and Pattern Recogni-* *tion (CVPR)*, 2017. 2

[35] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2

[36] N. N. Vo and A. F. Bobick. From stochastic grammar to bayes network: Probabilistic parsing of complex activity. In *CVPR*, 2014. 2

[37] J. Wu, Y. Yu, C. Huang, and K. Yu. Deep multiple instance learning for image classification and auto-annotation. In *Proceedings of the IEEE Conference on Computer Vision and* *Pattern Recognition*, pages 3460–3469, 2015. 2

[38] F. Xiao, L. Sigal, and Y. Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, 2017. 2

[39] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei. Every moment counts: Dense detailed labeling of actions in complex videos. *International Journal of* *Computer Vision*, 126(2-4):375–389, 2018. 1, 2

[40] W. Zhang, S. Zeng, D. Wang, and X. Xue. Weakly supervised semantic segmentation for social images. In *CVPR*, 2015. 2

[41] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *ICCV*, 2015. 2