

成 绩

****

**模式识别**

**实验报告**

|  |  |
| --- | --- |
| 院（系）名称 | 自动化科学与电气工程学院 |
| 专业名称 | 自动化 |
| 学生学号 | 21375174 |
| 学生姓名 | 冯沛林 魏干然 赵泽北  谢晓睿 湛梓昊 |
| 指导教师 | 秦曾昌 |

2024年11月

目录

[摘要 3](#_Toc188189996)

[1.介绍---湛梓昊 3](#_Toc188189997)

[2.相关工作---赵泽北 3](#_Toc188189998)

[3.预备---魏千然 3](#_Toc188189999)

[1. 时间动作分割（TAS）概述 3](#_Toc188190000)

[2. TAS定义 4](#_Toc188190001)

[3. TAS与其他任务 5](#_Toc188190002)

[数据集 6](#_Toc188190003)

[1、核心数据集 6](#_Toc188190004)

[2、数据集对比 8](#_Toc188190005)

[3、 数据集的时间动态（Temporal Dynamics）指标 9](#_Toc188190006)

[评价指标 10](#_Toc188190007)

[1、基于帧的度量 11](#_Toc188190008)

[2、基于段的度量 11](#_Toc188190009)

[3、无监督评估中的匈牙利匹配算法 12](#_Toc188190010)

[4模型介绍 14](#_Toc188190011)

[4.1Baseline Model: MS-TCN 14](#_Toc188190012)

[4.1.1 SS-TCN 14](#_Toc188190013)

[4.2 Base Model: MS-TCN2 16](#_Toc188190014)

[4.3 Imporved Model-1: MSTCN2+LSTM 17](#_Toc188190015)

[4.4 Imporved Model-2: MS-TCN2 18](#_Toc188190016)

[4.4 Imporved Model-3: MS-TCN2 19](#_Toc188190017)

[4.5 Imporved Model-4: Mamba 19](#_Toc188190018)

[5 实验设置 20](#_Toc188190019)

[5 实验结果—谢晓睿 21](#_Toc188190020)

[5.1 MSTCN+LSTM 21](#_Toc188190021)

[5.2 MSTCN+Global Attention 21](#_Toc188190022)

[5.3 MSTCN+Local Attention 22](#_Toc188190023)

[5.4 Mamba 23](#_Toc188190024)

[实验总结 23](#_Toc188190025)

[参考文献 24](#_Toc188190026)

# 摘要

# 1.介绍---湛梓昊

# 2.相关工作+排版和文献引用---赵泽北

# 3.预备---魏千然

### 时间动作分割（TAS）概述

时间动作分割（Temporal Action Segmentation, TAS）是一个视频理解任务，它在时间上对未修剪的视频序列进行分割。每个片段都用一组预定义的动作标签中的一个进行标记（如图1）。

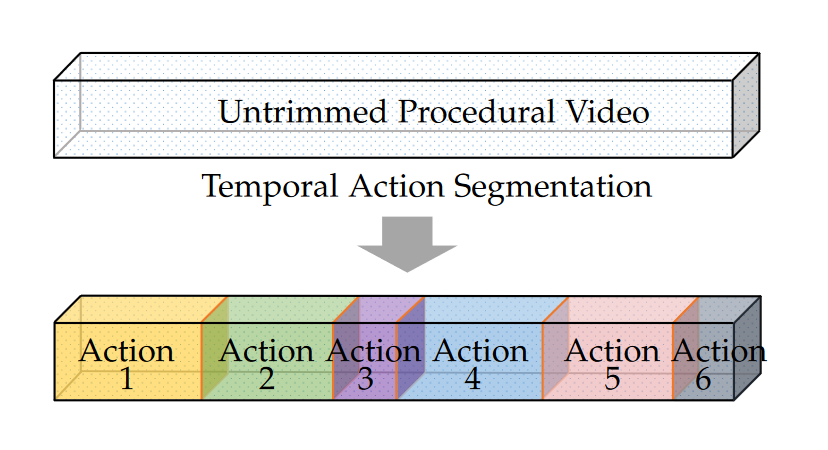


图 1

这项任务是更成熟的语义分割的一维时间类比，用帧级动作标签替换像素级语义标签。它能自动分割未修剪的视频序列，有助于理解正在执行的动作，比如动作何时开始，进展到什么程度，动作如何改变环境，以及预测人们接下来会做什么。它还支持各种下游应用，如辅助技术、视频监控和人机交互。

在计算机视觉中，动作识别是视频理解的标志性任务。在动作识别中，几秒钟的预剪辑视频被分类为单一语义标签。最先进的方法可以区分数百个类别。然而，对预剪辑视频进行分类的应用场景受限，因为监控系统、自动驾驶车辆和其他现实世界系统中的视频流是连续的，个别动作或事件是相关的，且可能远远超过几秒钟。因此，标准的动作识别方法并不直接适用。而动作分割方法主要针对未修剪的视频序列。这些视频描绘了一系列多个动作，通常跨越数分钟。比如一个常见的“制作咖啡”程序视频可能包括以下步骤：“拿杯子”、“倒咖啡”、“加糖”、“倒牛奶”和“搅拌咖啡”。描述整体过程的通用术语是“活动（activity）”，而组成步骤是“动作（actions）”。这些步骤通常遵循松散的时间顺序，即动作的某些排列（如“加水”和“磨咖啡”，见图2）。一个有效的分割模型应该使用序列信息来确定动作边界。

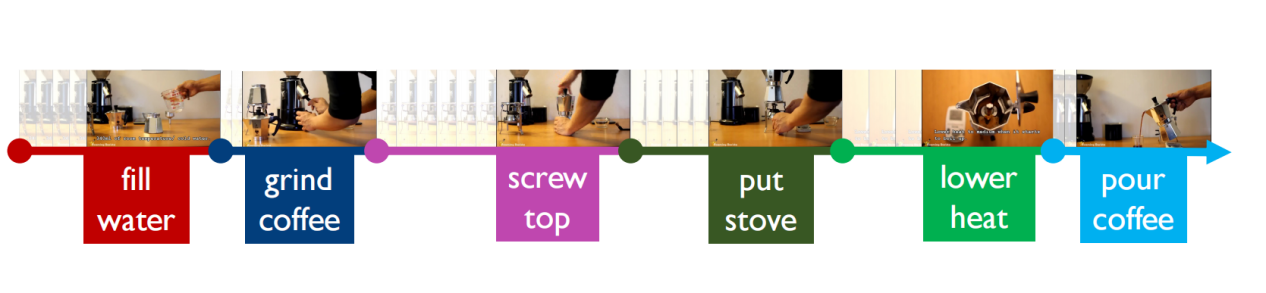


图 2

### TAS定义

时间动作分割（TAS）将未修剪的视频分割成时间段，并为每个片段分配预定义的动作标签。给定一个长度为T的视频 ，包含N个动作，分割方法产生以下输出：



其中 代表长度为 的视频片段，标签为 ，*C*是预定义的类别集合，并且任何 片段在时间上是连续的。这项任务也可以被视为语义（图像）分割的一维类比，并且可以被表述为逐帧动作分类，即:



其中*yt*是帧*t*的动作标签。方程（1）中的片段表述通常用于弱监督工作，这些工作预测最可能的动作序列，而方程（2）中的逐帧表述通常用于有密集标签的完全监督方法。这两种表述是等价的，可以从一种重构出另一种。

### TAS与其他任务

在视频理解领域，有一些与时间动作分割（TAS）密切相关的任务。他们之间可以根据数据域、段语义的识别和段之间的时间动态推理来区分。下面介绍这些相关的任务

1. 时间动作检测/定位（Temporal Action Detection/Localization, TAD/L）：这个任务检测动作实例的开始和结束，并同时预测语义标签。TAD/L处理日常生活中的普通视频，如THUMOS14，并从时间上稀疏的动作注释中学习。TAD与TAS的区别见图3所示：

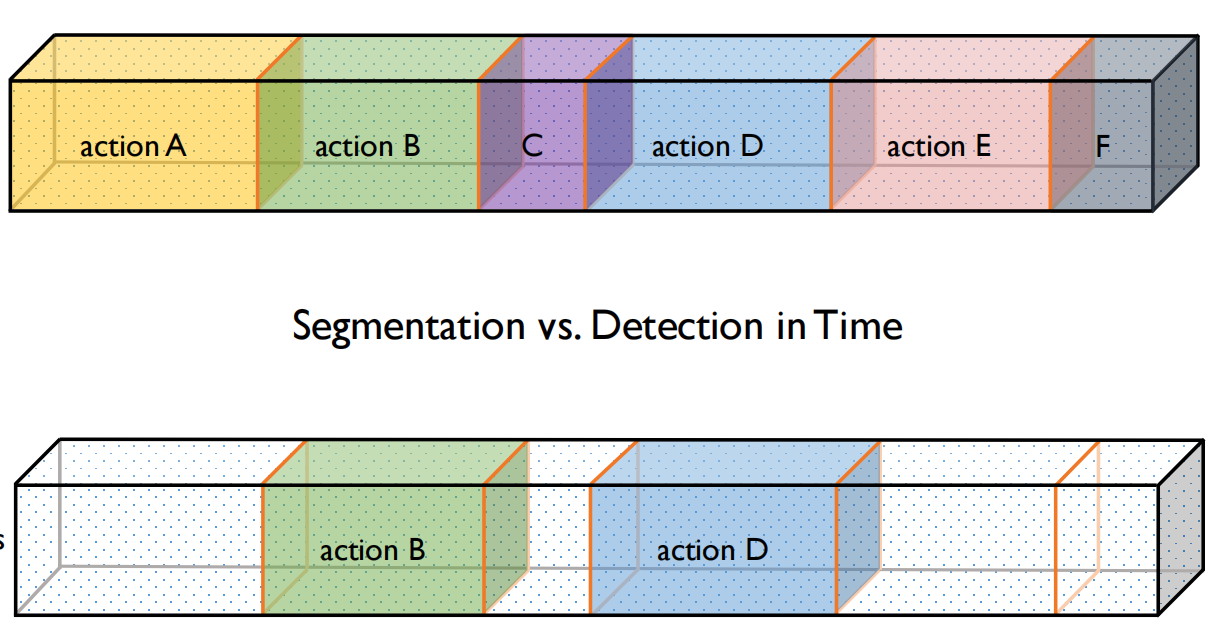


图 3

2. 序列分割（Sequence Segmentation, SS）：在其他领域，如运动捕捉数据和音频信号中应用广泛。大多数方法开发用于分割单个序列，而有些方法同时关注多个运动捕捉记录。

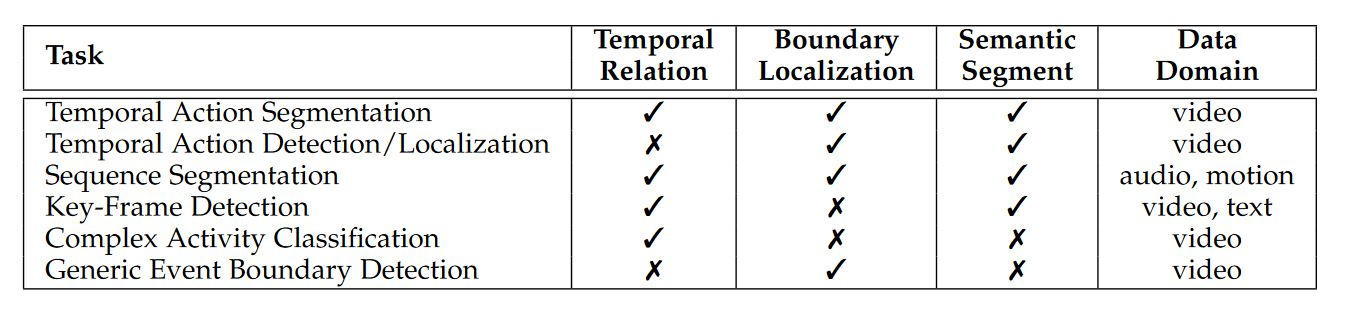
3. 关键帧检测（Key-Frame Detection, KFD）：识别动作的单个特征帧或关键步骤。与TAS一样，KFD需要对动作间的时间关系进行建模，但它的目标不是确定动作转换的边界。

4. 复杂活动分类（Complex Activity Classification, CAC）：在复杂活动层面上对程序性活动视频进行分类。这项任务应用于与TAS相同的数据，但不在帧级别分辨率上工作。

5. 通用事件边界检测（Generic Event Boundary Detection, GEBD）：识别视频中人类感知为类不可知事件边界的时刻，表明动作、主体和环境的变化。与TAS不同，GEBD不涉及语义标签，也不对检测到的边界之间的时间关系做假设。

这些任务与TAS有相同的特征，但每个任务都有其独特的重点和难点。下表展示了各任务之间的联系与区别。

表 1



## 数据集

### 1、核心数据集

在时间动作分割（TAS）任务中，有一些常用的核心数据集，记录了人们执行程序性活动，例如准备餐点或组装家具。这些数据集都标注了动作片段的开始、结束和类别标签，用于训练模型。它们在数据大小、领域、视角等方面各有不同，并在时间动态、动作持续时间分布上有差异。这里介绍几种常见数据集：

1. GTEA (Georgia Tech Egocentric Activities) ：

特点：包含28个视频，记录了4名参与者在单一厨房环境中进行的7种烹饪活动，视频由参与者佩戴的帽子上的摄像机录制，提供第一人称视角。

数据规模：0.4小时的视频时长，0.5K个动作片段，71个动作类别。



2. 50Salads：

特点：包含50个视频，展示了两种混合沙拉的制作过程，视频从顶部视角捕捉，参与者遵循从统计食谱模型中随机选择的步骤。

数据规模：5.5小时的视频时长，0.9K个动作片段，17个动作类别。



3. Breakfast：

特点：包含1712个视频，记录了52名参与者在不同厨房环境中进行的10种早餐相关活动，视频从第三人称视角录制。

数据规模：77小时的视频时长，11K个动作片段，48个动作类别。



4. YouTube Instructional ：

特点：包含150个视频，涵盖5种教学活动，每个活动有30个视频，主要用于无监督的时间动作分割任务。

数据规模：7小时的视频时长，5个动作类别。



5. Assembly101：

特点：包含4321个视频，记录了53名参与者在没有指导的情况下拆卸和重新组装可拆卸玩具的过程，提供手-物交互的细粒度注释和粗略的动作标签。

数据规模：513小时的视频时长，1M个动作片段，202个动作类别。

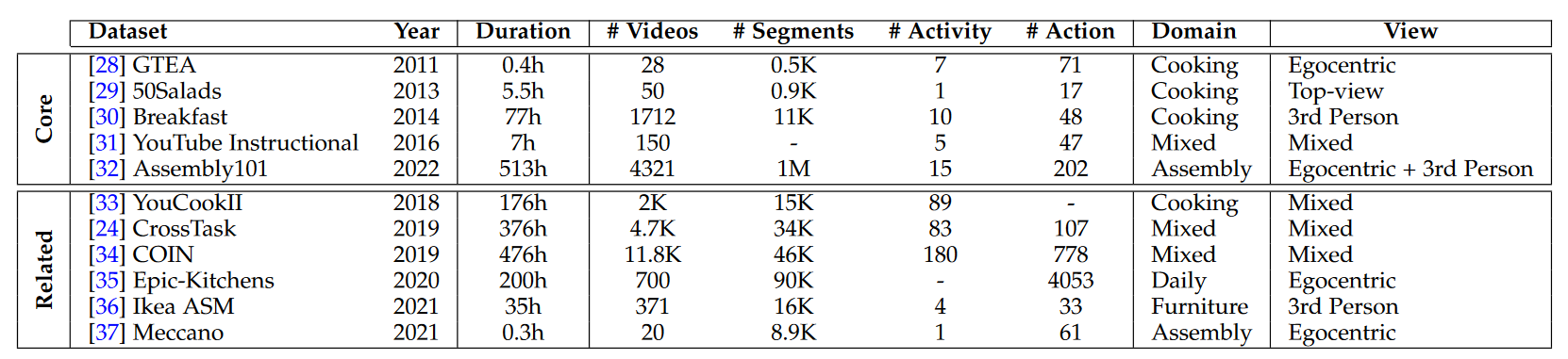


本文用到的数据集是50Salads、Breakfast、GTEA

### 2、数据集对比

表2对各类程序视频数据集进行了比较，并根据它们的来源、规模、动作数量和视角进行了分类。

表 2



这些数据集通常采用静态的第三人称视角或动态的第一人称视角进行录制。第三人称视角的静态背景有助于在同一视频内识别不同的动作，而背景的变化则增加了跨视频识别动作的难度。第一人称视角在捕捉物体和工具方面表现优异，有助于识别手部与物体的交互，这对程序性活动的理解至关重要。然而，第一人称视角下的相机运动也带来了额外的挑战。例如，Epic-Kitchens数据集就是一个大规模的、以第一人称视角记录厨房活动的显著例子。也有一些小规模的、关注烹饪活动的第一人称视角数据集。相比之下，Breakfast数据集主要包含了从多个第三人称视角录制的视频。在这些数据集中，只有Assembly101提供了第一人称和第三人称视角的同步视图。然而，现有的程序活动数据集在多样性上相对有限，大多数数据集如50Salads和Breakfast专注于烹饪和厨房活动。目前，Assembly101是唯一一个非烹饪领域的大规模数据集。

### 数据集的时间动态（Temporal Dynamics）指标

动作分割任务的一个定义特征是动作之间的时间动态。为了定性评估时间动态，我们定义了重复分数（repetition score）和顺序变化分数（order variation score）。序列中动作重复的程度通过重复分数*r*来量化，定义为 :



其中u是序列中动作的种类，g是序列中动作的总数。分数r的范围在0到1之间，其中0表示没有重复，更高的分数表示序列内重复的程度更大。具体示例如下图所示：



图 4

左图中共有7个动作，属于4个种类，故重复分数为 ；右图中共有7个动作，属于7个种类，故重复分数为 。

顺序变化分数*v*定义为1减去两个序列 之间平均编辑距离 根据最大序列长度的归一化。



分数v的范围也是[0, 1]，其中1表示动作对之间的顺序没有偏差。相反，较低的分数表示有更多的顺序变化。具体示例如下图所示：

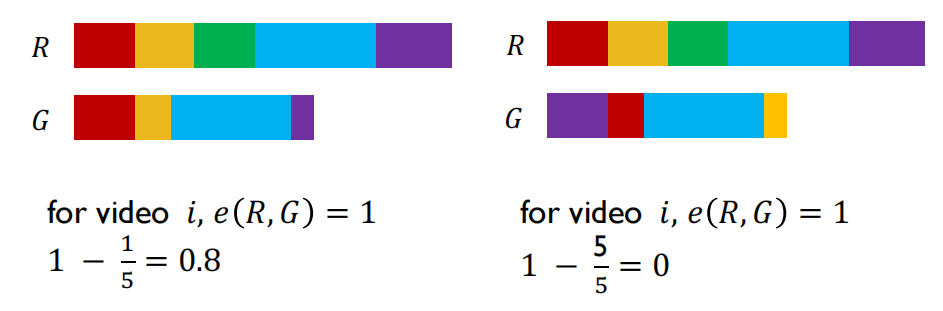


图 5

左图两序列的平均编辑距离为1，最大序列长度为5，故顺序变化分数

右图两序列的平均编辑距离为5，最大序列长度为5，故顺序变化分数

Assembly101将自己定位为动作之间序列动态建模的挑战性基准。正如表3所示，Assembly101的顺序变化分数介于Breakfast和50Salads之间，并且相对于这两个数据集，它包含了相对更多的重复步骤。

表 3



## 评价指标

在时间动作分割（TAS）中，三个常用的评估指标是帧平均准确率（Mean over Frames, MoF）、编辑分数（Edit Score）和F1分数（F1 Score）。第一个是基于帧的度量，而后两者是基于片段的度量。所有这三个指标在全监督、弱监督和半监督学习中都有使用。对于无监督学习，可以采用匈牙利匹配算法来进行评估。

### 1、基于帧的度量

帧平均准确率（MoF）是帧级别的准确率，定义为模型正确帧预测的比例：



MoF指标在数据集不平衡的情况下可能会有问题，即如果频繁和长时间的动作类别占主导地位。当前数据集的长尾特性意味着具有相似MoF分数的模型可能存在很大的定性差异，这表明采用类别平衡的指标可能更合适。此外，作为逐帧计算的MoF没有捕捉到片段质量；即使在片段被分割成许多不连续的子片段时，分数也可能很高。将一个动作分成许多不连续的子片段被称为过度分割。像编辑分数和F1分数这样的基于片段的度量更适合评估过度分割等现象。

### 2、基于段的度量

编辑分数量化了两个序列的相似性。基于Levenshtein或编辑距离，计算将一个片段序列转换为另一个片段序列所需的最小插入、删除和替换操作的数量。考虑X和Y作为预测和真实动作片段的有序列表，累积距离值e定义为：



其中 ,  分别是X和Y的索引， 是指示函数。方程(6)可以通过动态规划有效解决。最终的编辑距离值然后通过两个序列的最大长度进行归一化，以计算编辑分数：



作为一个度量，编辑分数可以评估模型预测动作序列的能力，而不需要与真实动作之间的精确帧级对应关系。

F1分数（或F1@τ）比较每个片段与相应真实动作的交集比（Intersection over Union, IoU），基于某个阈值τ/100。如果一个片段与真实动作的IoU超过阈值，则被视为真阳性。如果在单个真实动作的跨度内有一个以上的正确片段，则只有一个片段被视为真阳性，其他被视为假阳性。基于真阳性和假阳性以及假阴性（未检测到的片段），可以计算精确度和召回率，并将两者融合为F1分数：



常用的τ值是{10, 25, 50}。图6可视化了这三个评估指标。

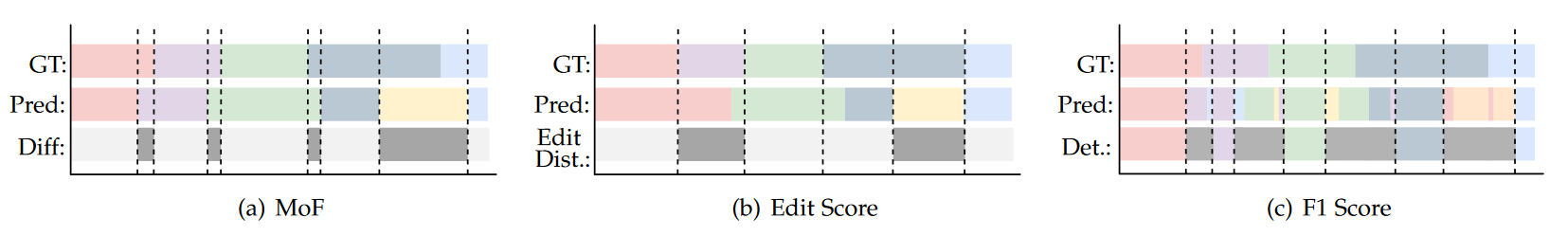


图 6

### 3、无监督评估中的匈牙利匹配算法

在无监督评估场景中，如果没有建立起预测片段与真实动作之间的对应关系，那么上述提到的评估指标就无法直接应用。匈牙利匹配算法是一个组合优化算法，它用于在二分图中寻找最大权重匹配，这一算法已被广泛用于评估无监督聚类任务。在无监督的时间动作分割（TAS）中，匈牙利匹配算法将给定的N个聚类框架X与M个动作标签集合Y进行最佳匹配 ：



这里， 表示属于第n个聚类的帧， 表示带有动作标签m的帧。 是分配给一对(n, m)的指示函数， 表示在聚类n中带有真实类别标签m的帧数。当两个集合的类别数量相等（N = M）时，匈牙利匹配算法构建的是一一对应关系。如果不等，则产生一个大小为min(N, M)的单侧完美匹配。未匹配的剩余聚类自动被视为背景。评估结果基于这些对应关系得出。根据二分子集的范围，匈牙利匹配算法可以在三个不同级别上应用，如图7所示。

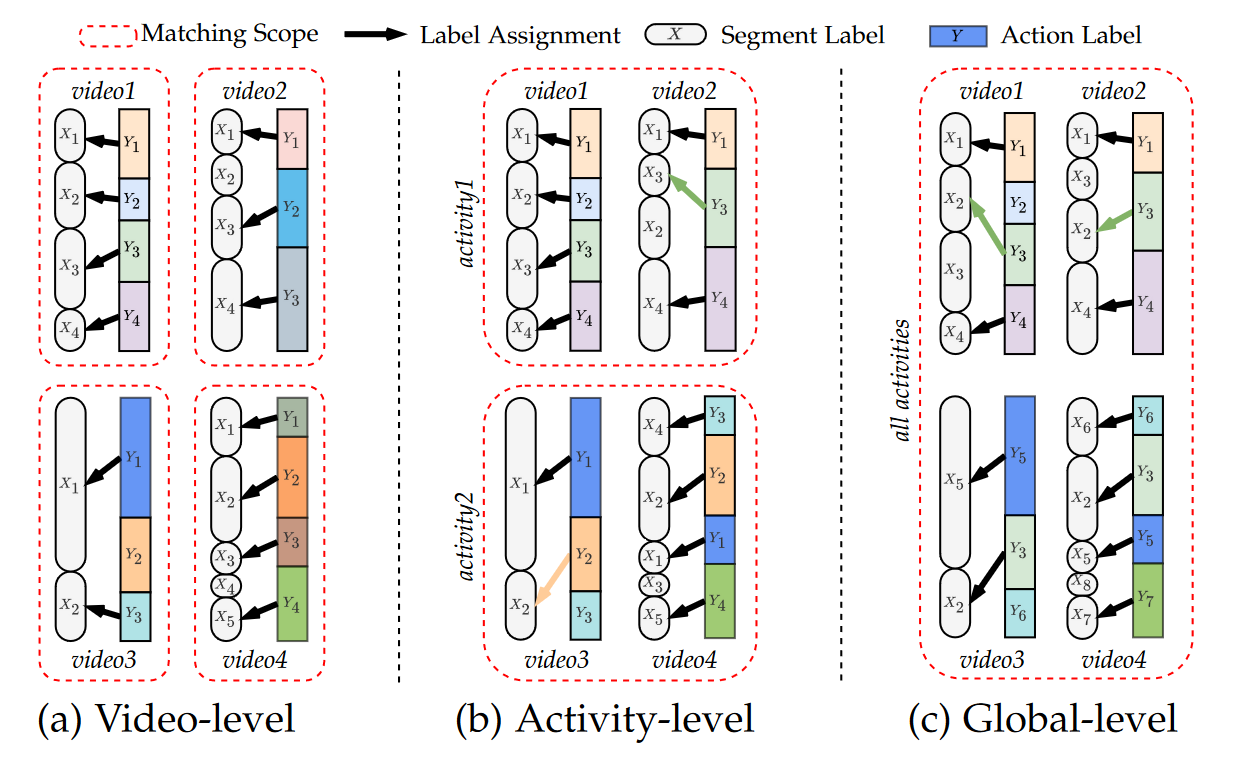


图 7

视频级匹配将聚类与单个视频的真实动作相对应。这种匹配评估了模型将视频序列分割成不同动作的能力，并因为范围有限而产生最高的评估分数。在每个匹配范围内，匹配对跨视频的关联不敏感。活动级匹配在每个复杂活动中将聚类与标签关联起来。大多数无监督工作遵循这种匹配范围，即处理来自同一活动的视频。活动级别的分组会导致分配变化，如图7(b)中的彩色箭头所示。全局级匹配在整个数据集上执行，是最具有挑战性的设置，因为它需要同时考虑活动内和活动间的匹配。匈牙利匹配的不同范围对应于TAS模型的不同学习目标；范围越大，任务越通用。视频级匹配要求模型在同一视频中区分不同的动作，即视频内动作的区分。对于活动级匹配，模型必须在视频中区分动作并形成活动内动作的关联。在全局级匹配中，模型必须包括活动间的关联，以构建跨复杂活动的可行动作对应关系。需要注意的是，在更广泛范围内学习到的模型可以向下兼容，可以调整为在更细的范围内进行评估，例如从全局到活动级别，但反过来则不行。尽管这样做在实践中是可行的，但由于模型的不同学习要求，结果并不直接可比。

# 4模型介绍

## 4.1Baseline Model: MS-TCN

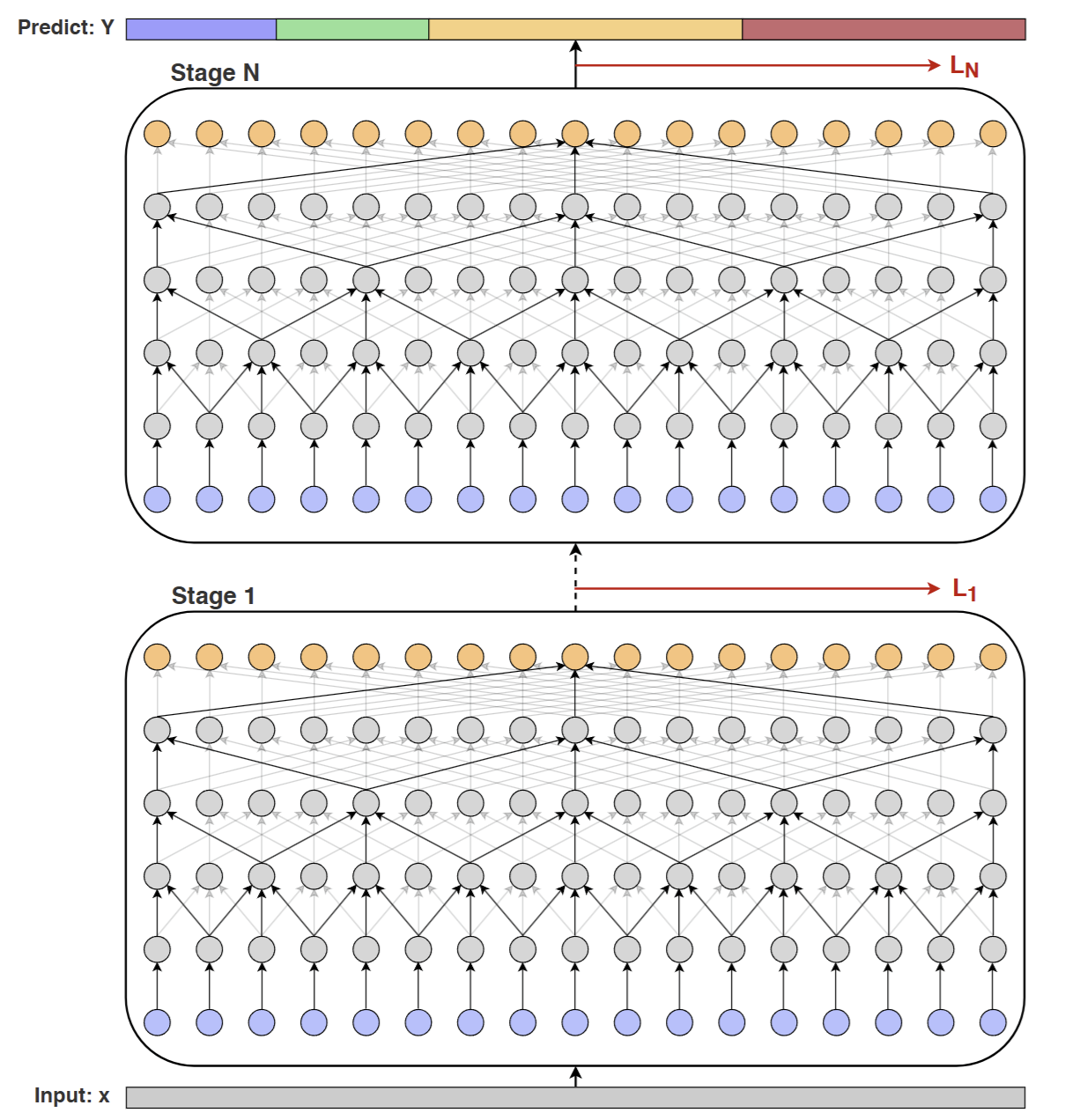


图1：MS-TCN网络结构

MS-TCN由Single-StageTCN（SS-TCN）和Multi-Stage TCN两个架构组成

## 4.1.1 SS-TCN

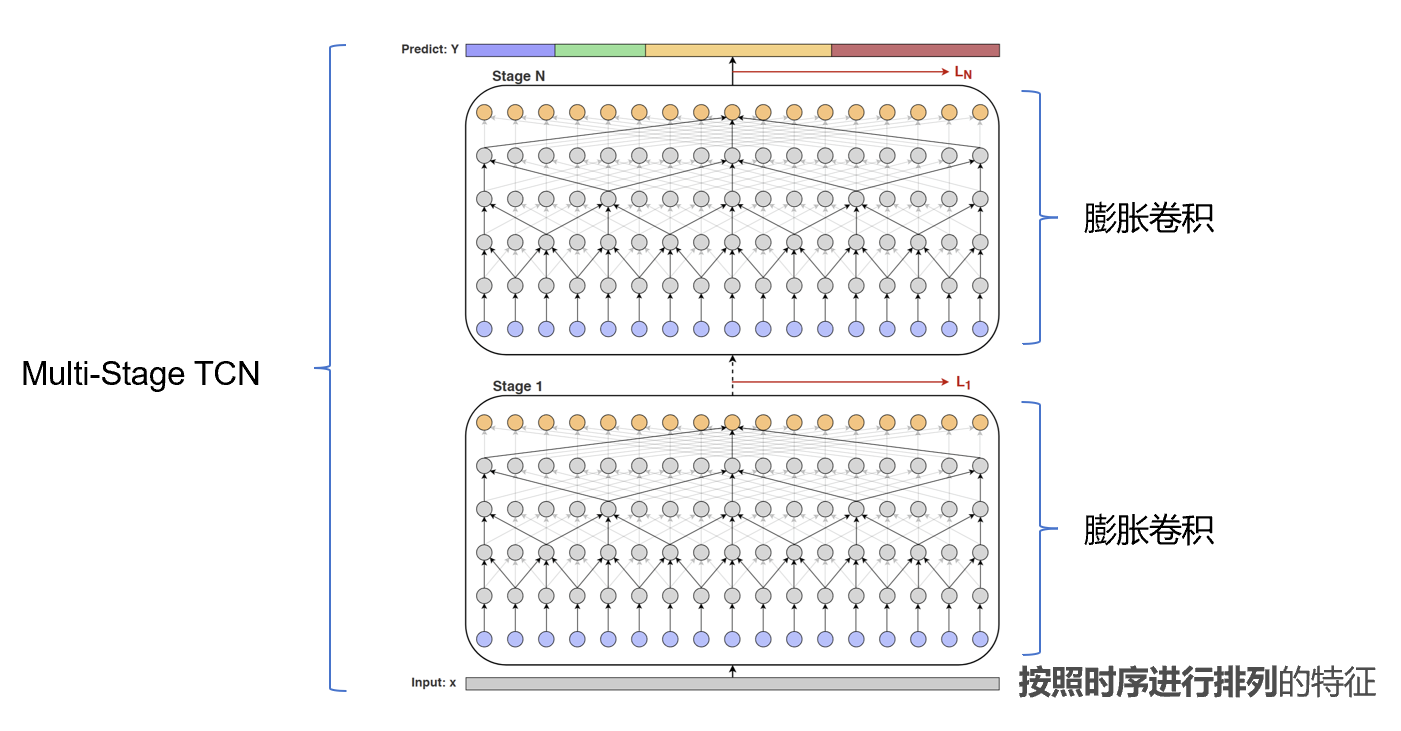
SS-TCN 的目的是在给定输入序列的情况下，捕获序列中的时间依赖性和特征模式。通过多个堆叠的时域卷积层，SS-TCN 能够高效地建模长时间范围的依赖性，同时保证计算效率。

第一层SS-TCN是一个1\*1的卷积层，该层的作用是调整输入特征的尺寸来匹配在网络中的feature-map。下一层是一些1D的膨胀卷积（Dilated Convolution）,膨胀卷积的作用是捕获序列中远距离的时间依赖性，其操作类似于池化，还可以增加增加卷积核的感受野。SS-TCN中还有Relu激活函数增加非线性拟合能力，以及残差结构减小梯度消失问题。

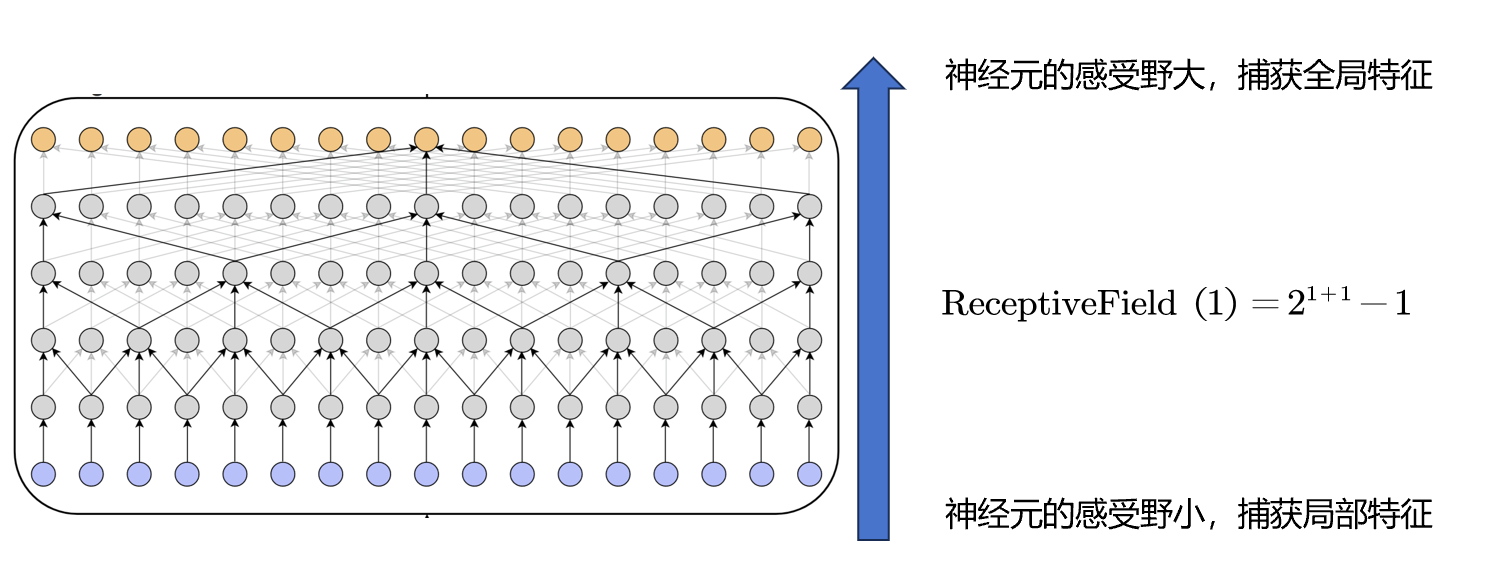
4.1.2 MS-TCN

在这个多阶段模型中，第一阶段的输入是视频的逐帧特征，接下来每个阶段都从前一阶段进行初步预测，并在后一阶段进一步预测。这种设计的motivation在于采用逐步优化预测策略：初始阶段可能存在错误预测，后续阶段能够多次利用上下文信息修正这些错误。另一方面动作分割中的一个难点是精确预测动作边界，多个阶段能够更细致地调整边界位置，达到细化边界的效果。

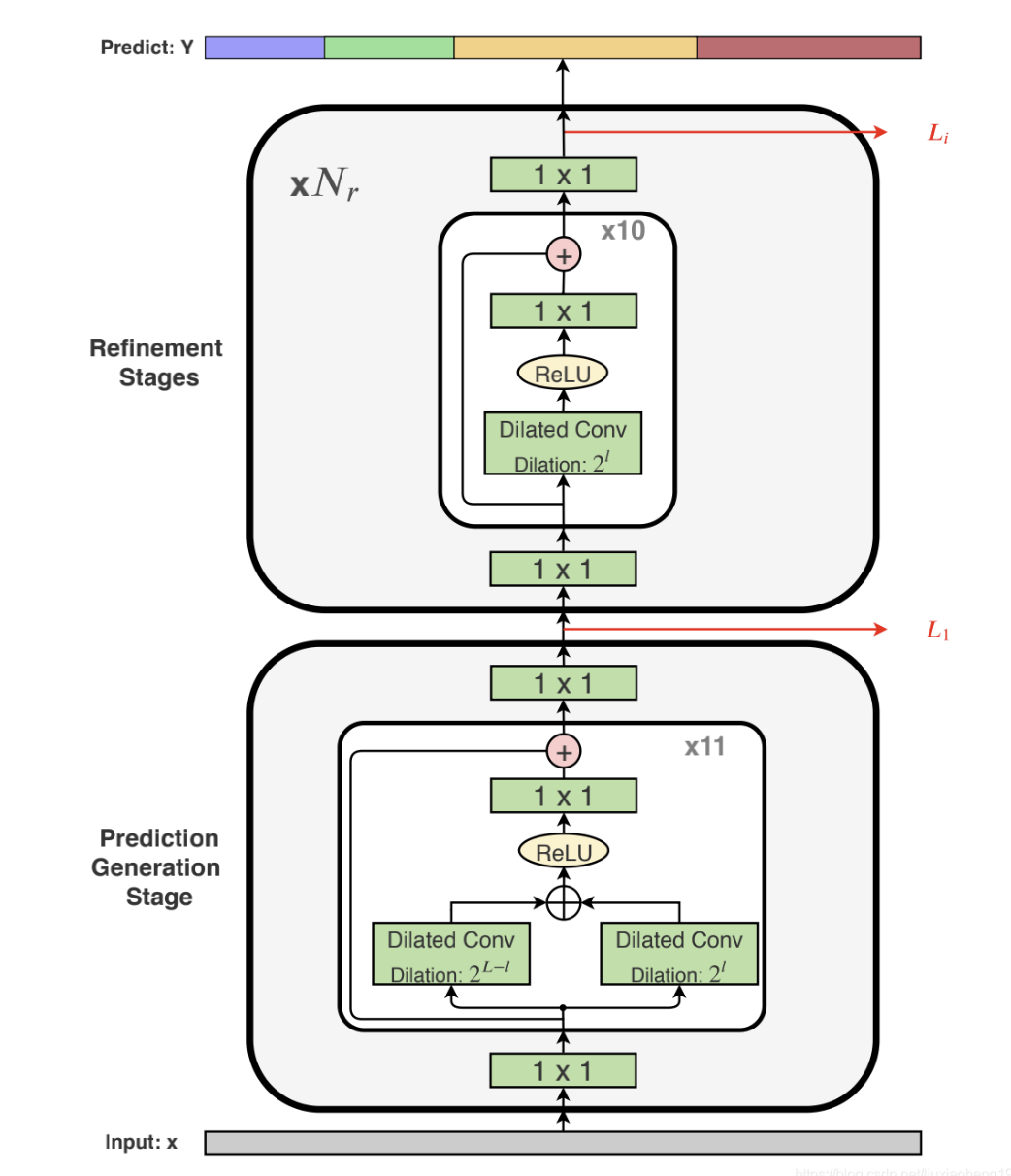
MS-TCN的架构如下图所示：



其主要是由卷积核感受野呈幂指数变化的膨胀卷积神经网络组成。随着网络层数的加深，神经元的感受野逐渐变大，从刚开始捕捉局部特征逐渐转化为捕获全局特征。

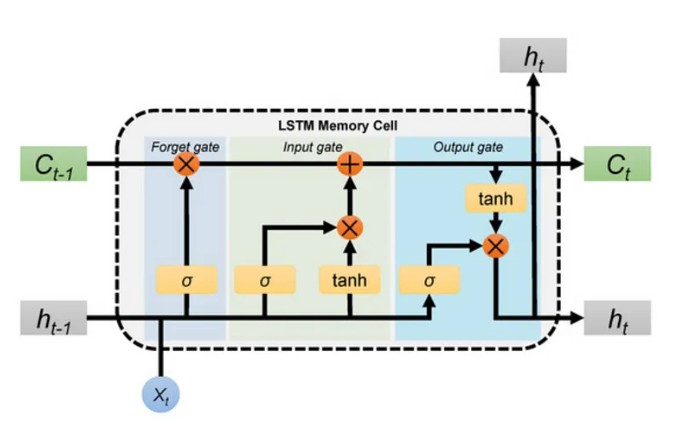


## 4.2 Base Model: MS-TCN2



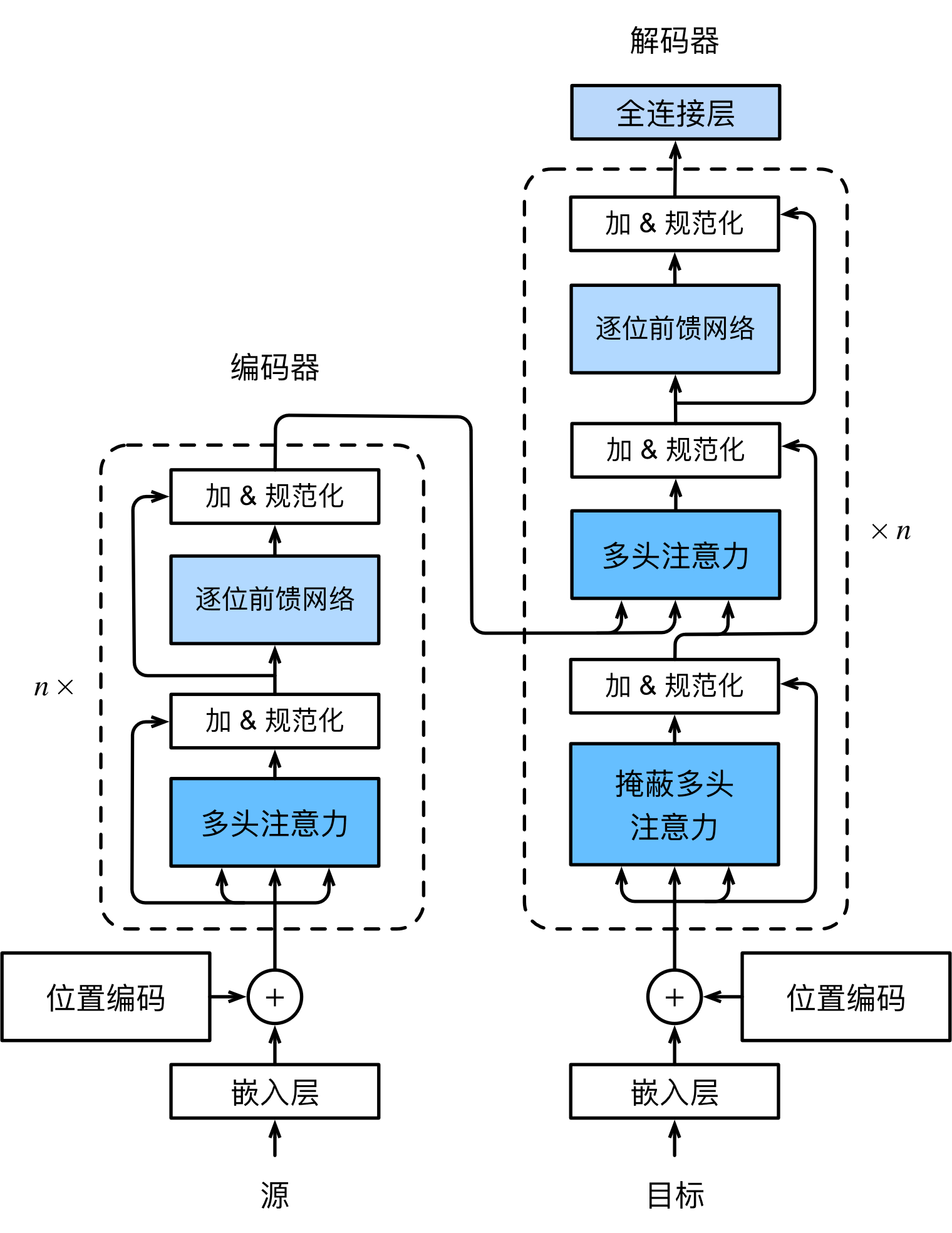
MS-TCN2只是在Prediction Generation部分进行了改进，另外的Refinement Stages保持不变。在Prediction Generation中，引入了感受野逐渐就爱你喜爱哦的收缩卷积分支，来达到和膨胀卷积的平衡。

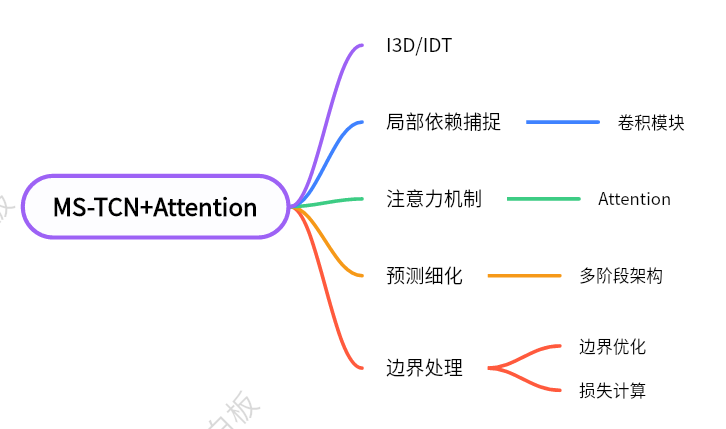
## 4.3 Imporved Model-1: MSTCN2+LSTM



MS-TCN++主要依赖多阶段的卷积层捕捉时间依赖，其感受野虽然较大，但仍可能对长时间跨度的复杂动作或跨阶段的动作序列建模不足。为了解决这个问题，我们引入了LSTM增强MS-TCN的场时间跨度的序列建模能力。

## 4.4 Imporved Model-2: MS-TCN2

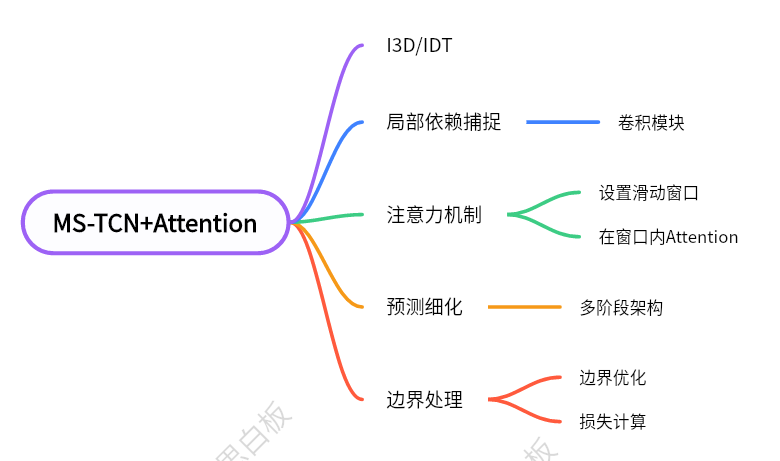




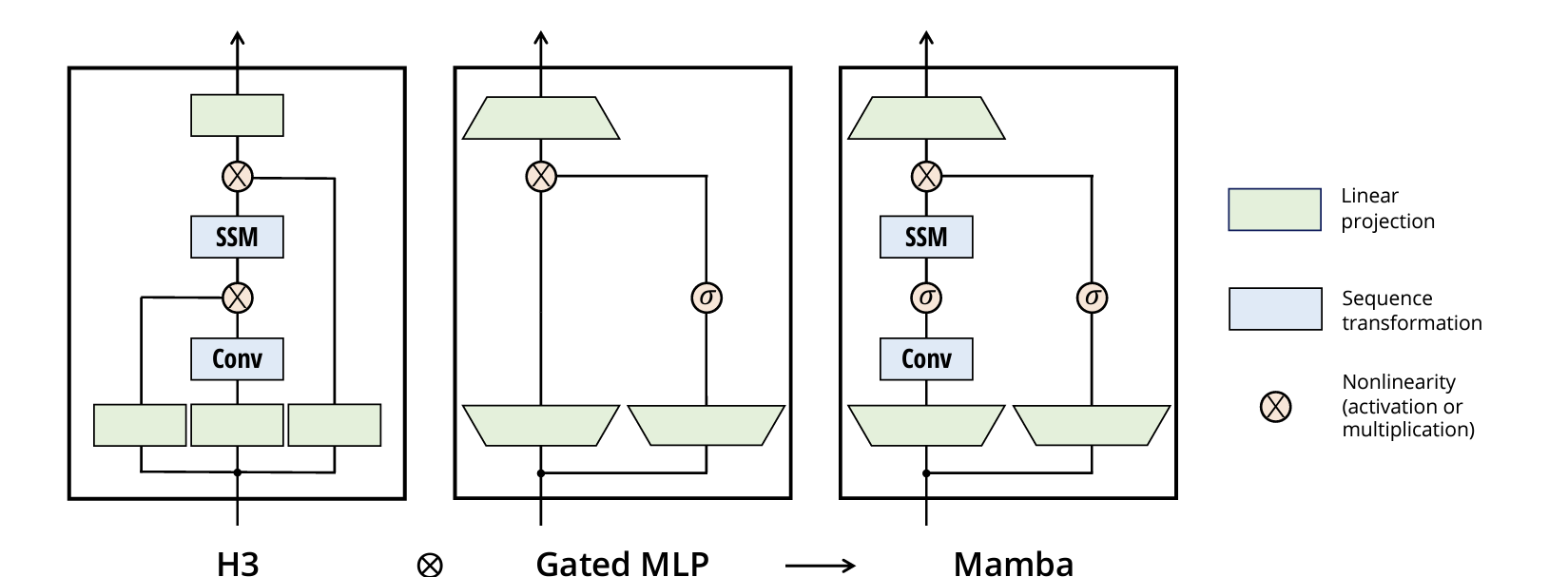
Transtormer架构近年来在时间序列和视频建模中表现出了卓越的能力，其关键特性在干多头自沣意力机制能够灵活捕捉时间步之间的长短期依赖。将Transformer与MS-TCN++结合，可以显著增强模型处理非线性时间依赖的能力。

## 4.4 Imporved Model-3: MS-TCN2

由于I3D提取的序列长度大于2k，自注意力模块直接计算全局注意力计算量较大并且捕获能力不佳。因此我们将注意力改进为了windowsize=50的滑动窗口注意力增强注意力模块的捕获能力。新的模型架构如下图所示：



## 4.5 Imporved Model-4: Mamba



Mamba架构是一种新型的深度学习模型，旨在解决传统Transformer在处理长序列数据时计算复杂度高、效率低的问题。与Transformer相比，Mamba在吞吐量上提高了约5倍，且在性能上与两倍规模的Transformer相当.

Mamba架构的核心思想：1) 选择性状态空间模型（Selective State Space Models，SSM）：Mamba基于SSM，通过引入选择性机制，能够有效捕捉长程依赖关系。与传统的循环神经网络（RNN）相比，SSM具有线性时间复杂度，避免了RNN在处理长序列时的梯度消失或爆炸问题。2) 硬件感知计算：Mamba设计时考虑了硬件的并行计算能力，采用了并行关联扫描（Parallel Associative Scan）和内存重新计算（Memory Recomputation）等技术，提升了训练和推理的效率。

3) 模块化设计：Mamba采用模块化设计，将SSM与门控多层感知机（Gated MLP）相结合，形成新的计算单元。这种设计使得模型在处理序列数据时，既能捕捉长程依赖，又能保持计算效率。

4) 线性时间复杂度：Mamba的设计使其在处理长序列时，计算复杂度保持在O(n)，避免了Transformer在处理长序列时计算量呈平方级增长的问题。

我们将Prediction Generation的膨胀卷积改为了四层的Mamba架构。

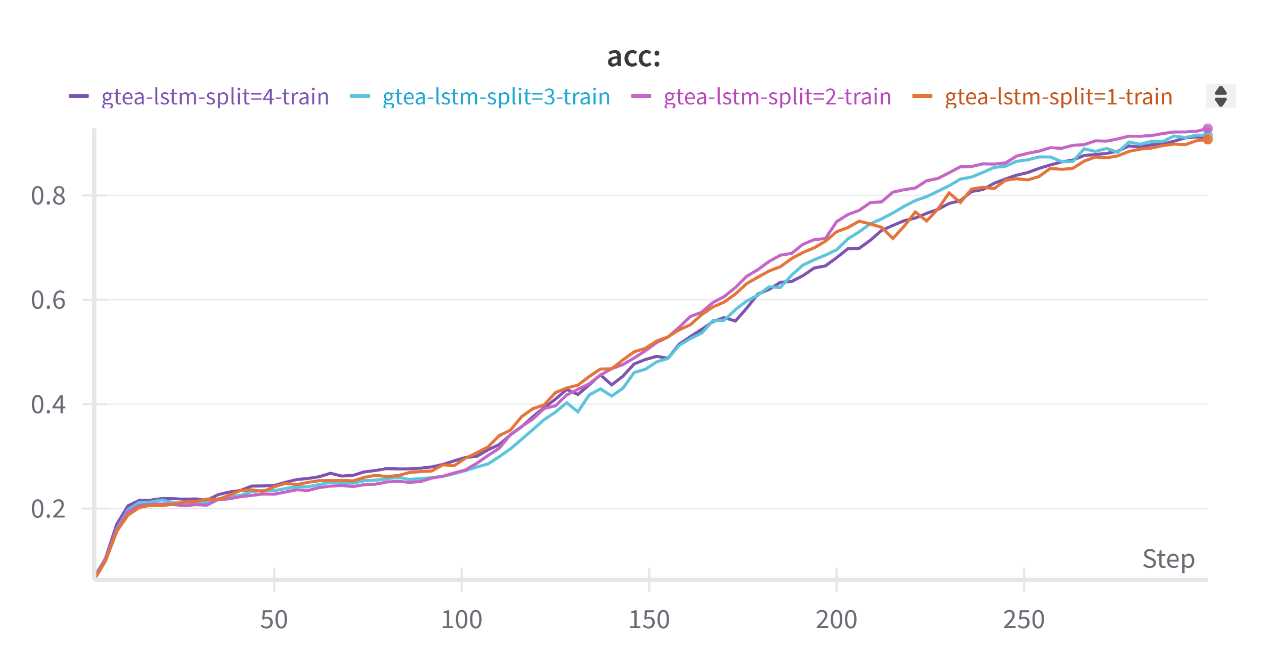
# 5 实验设置

|  |  |
| --- | --- |
| Loss Function |  |
| 特征提取 | I3D (Inflated 3D ConvNet) |
| num\_stages | 4 |
| num\_layers | 11 |
| lr | 0.0005 |
| batch\_size | 4 |
| num\_epochs | 50 |
| Optimizer | Adam |

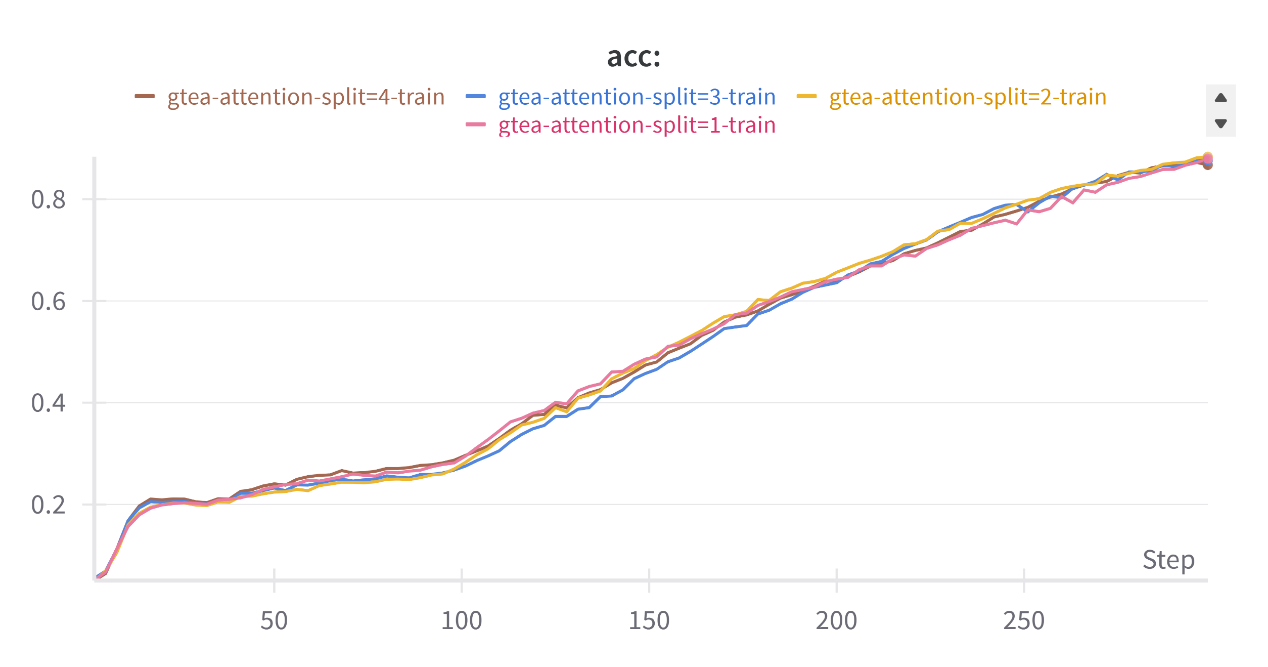
由于算力有限，我们在Temporal Action Segmentation中最小的gtea数据集上进行了广泛的探索。

# 5 实验结果—谢晓睿

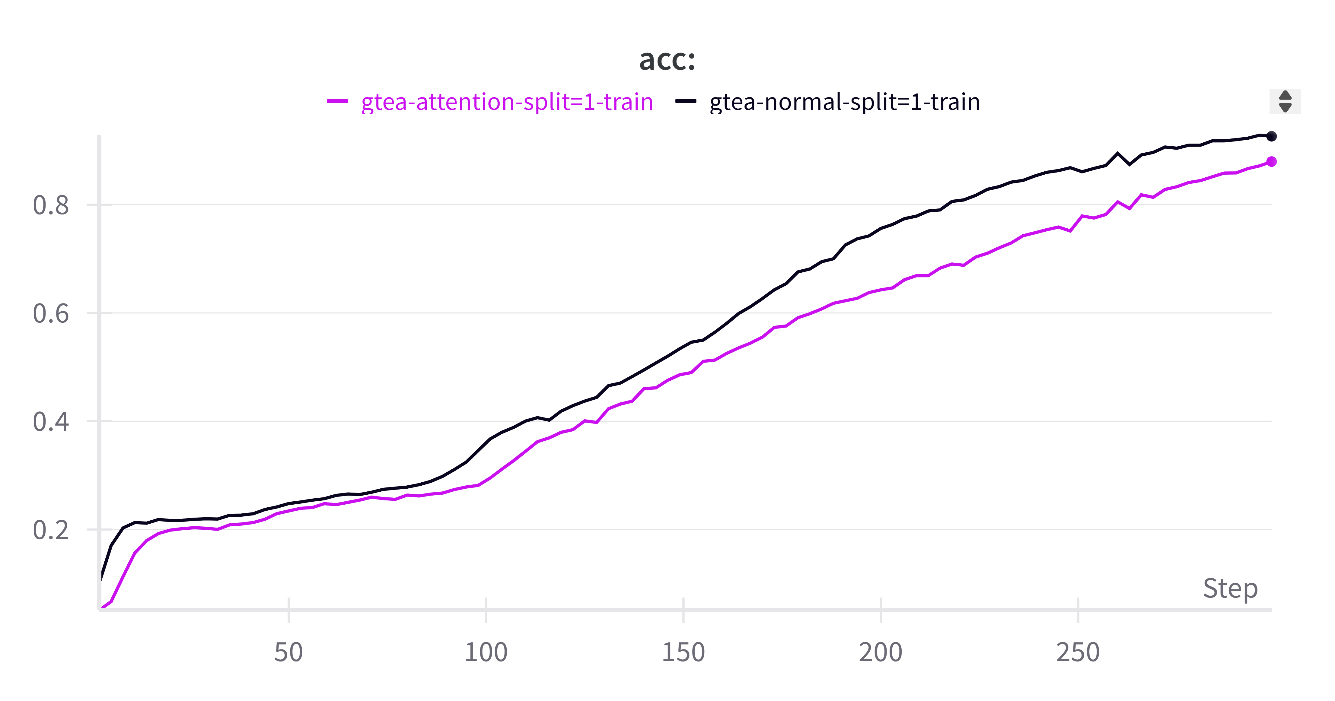
## 5.1 MSTCN+LSTM

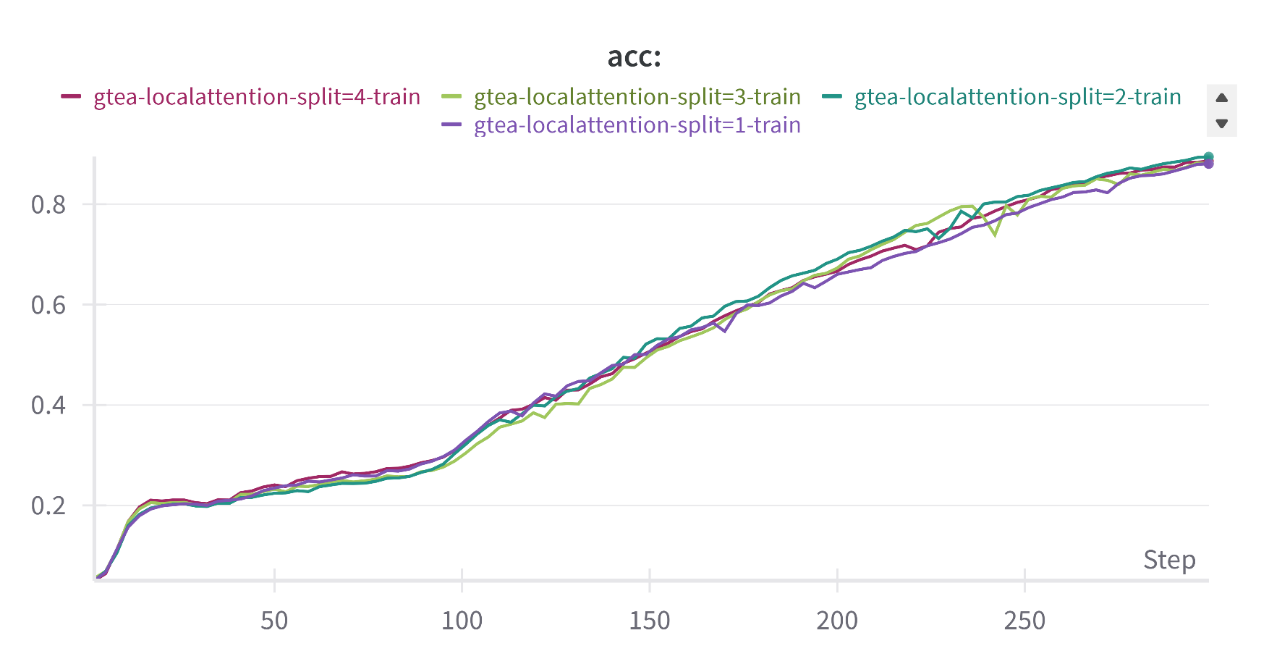


## 5.2 MSTCN+Global Attention

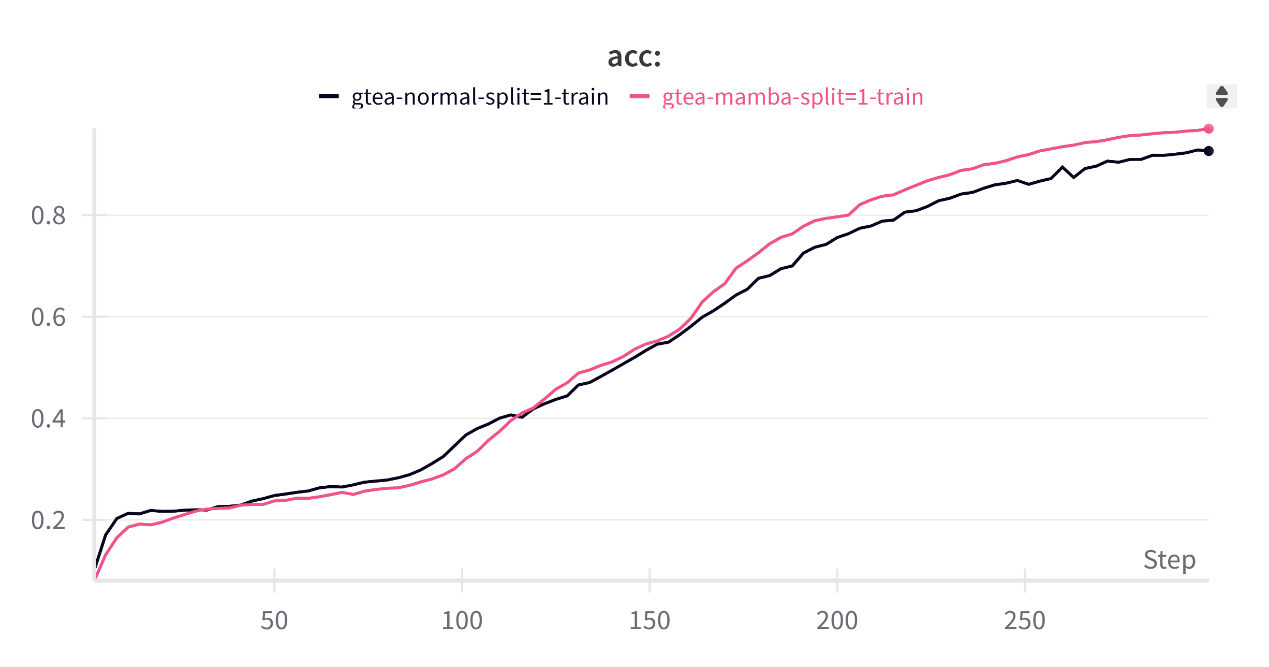


## 5.3 MSTCN+Local Attention



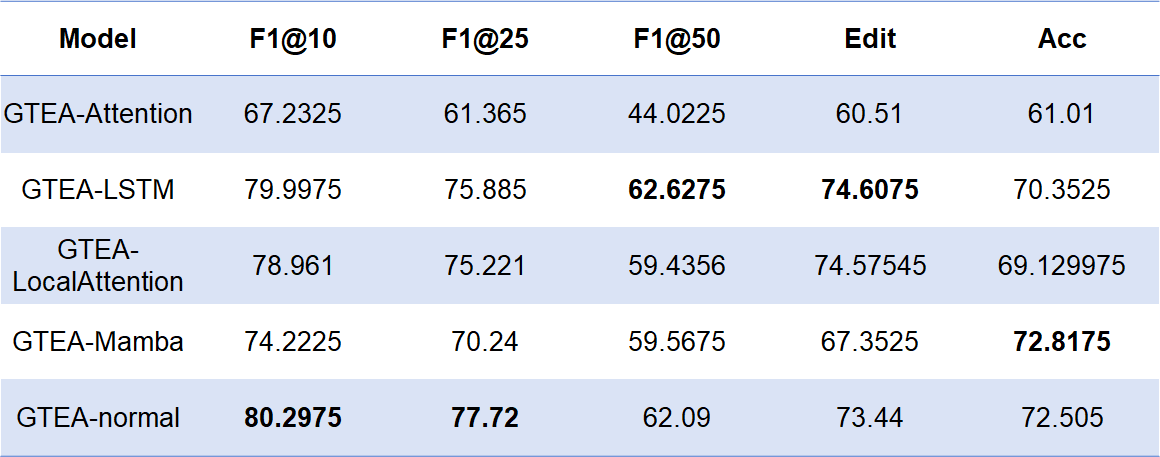


## 5.4 Mamba





# 实验总结



# 参考文献