ISCAS 中国科学院软件研究所学术年会’2022
暨计算机科学国家重点实验室开放周
关键技术

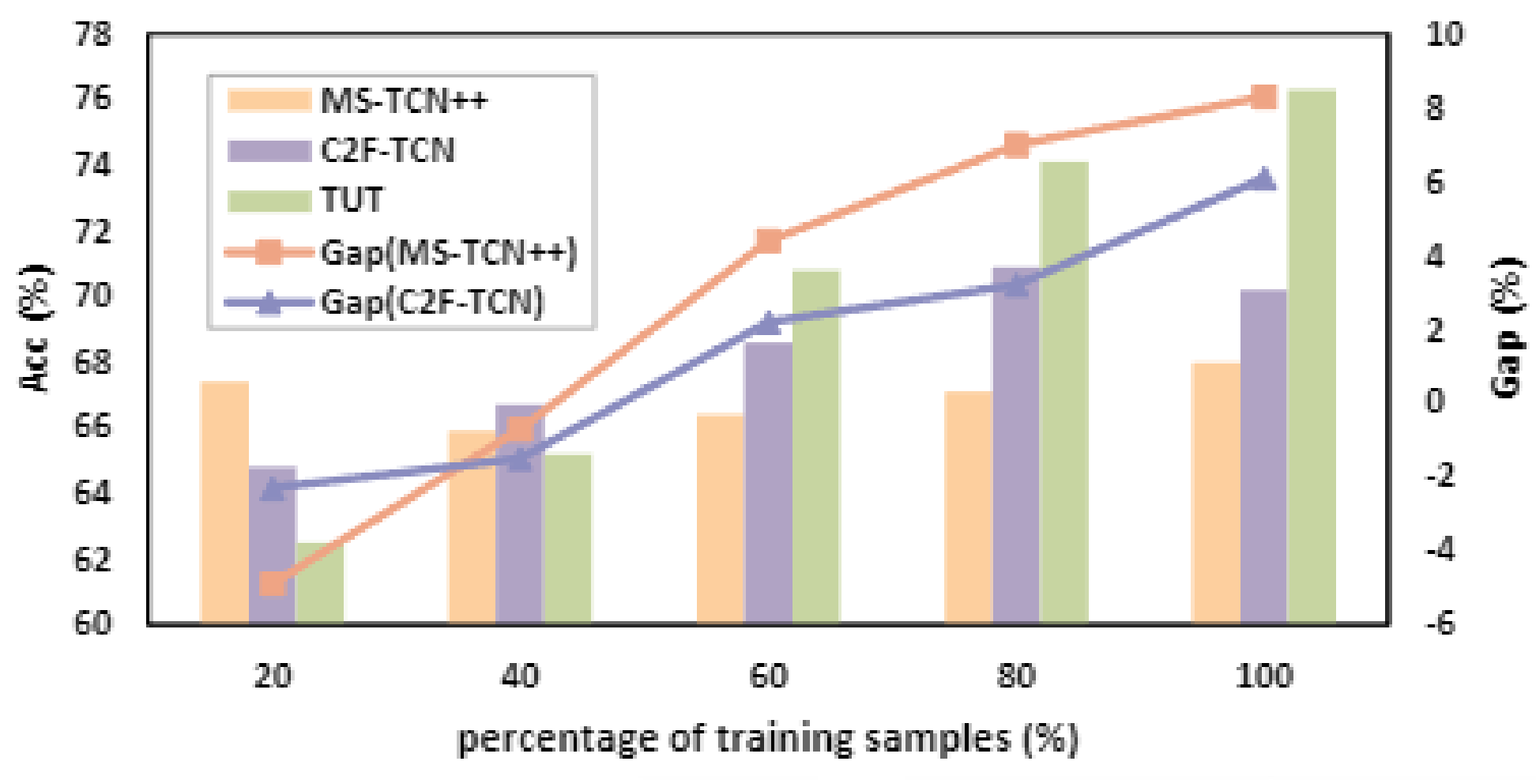
用于视频动作分割的Transformer模型

杜大钊，司凌宇

杜大钊 18800132373 dudazhao20@mails.ucas.ac.cn

1. 技术背景

由于互联网上视频数据的增加，视频理解已经发展成一个十分广阔的学术研究和产业应用方向。动作分类已经取得了很大的进展，但从长视频中分割和识别动作仍然是个具有挑战性的问题。



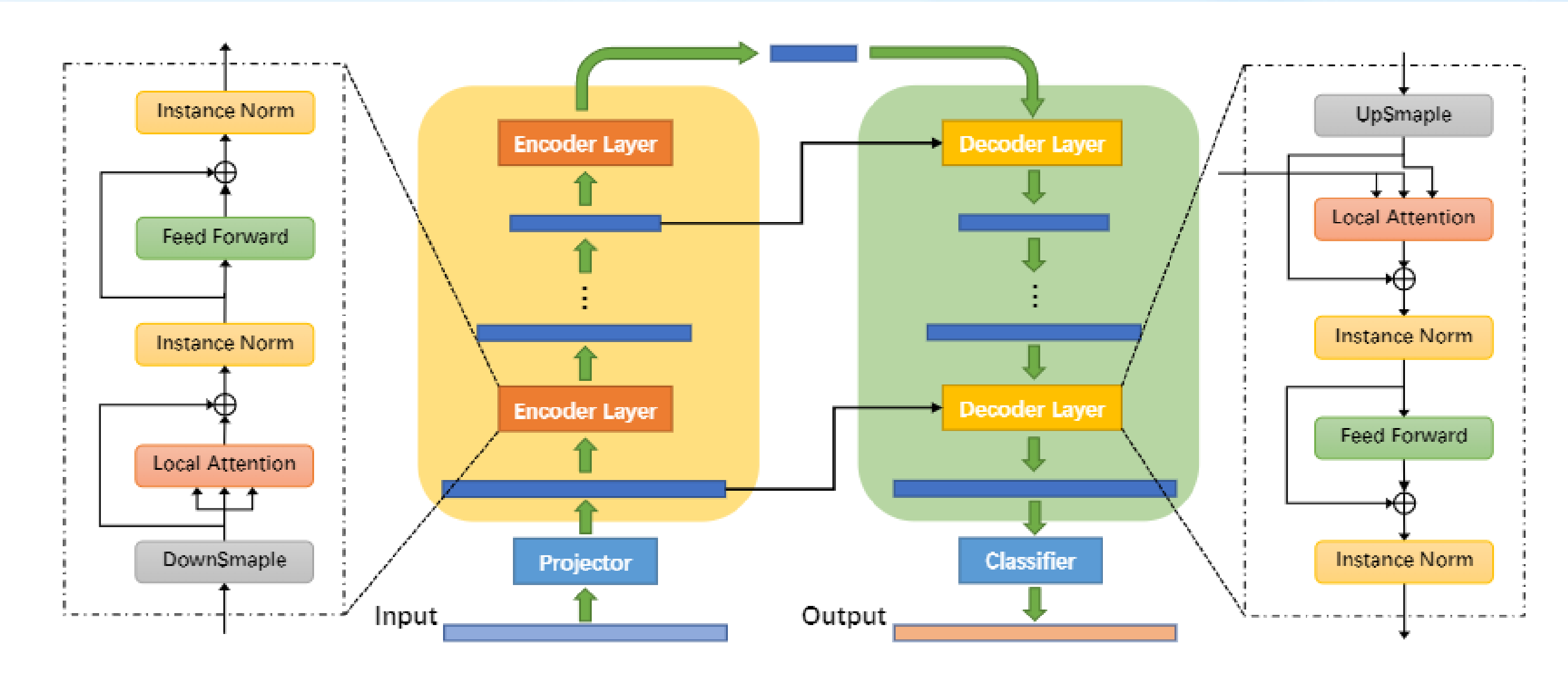
不同模型在各种训练数据量下的性能比较

大多数最先进的方法都专注于设计基于时间卷积的模型，如TCN，但时间卷积的不灵活性和建模长期时间依赖性的困难限制了这些模型的潜力。具有适应性和序列建模能力的基于Transformer的模型最近被用于各种任务。然而，由于缺乏归纳偏置和长视频序列处理效率低下，限制了Transformer在动作分割中的应用。

因此，通过加入时间采样，本技术设计了一个无时间卷积的基于的模型。它不仅降低了Transformer的复杂度，同时引入了归纳偏置，即相邻帧更有可能属于同一类，但引入粗分辨率会导致边界的错误分类。我们观察到边界帧与其相邻帧之间的相似性分布取决于边界帧是动作段的开始还是结束。因此，我们进一步提出了一种基于注意模块帧间相似性分数分布的边界感知损失，以增强识别边界的能力。

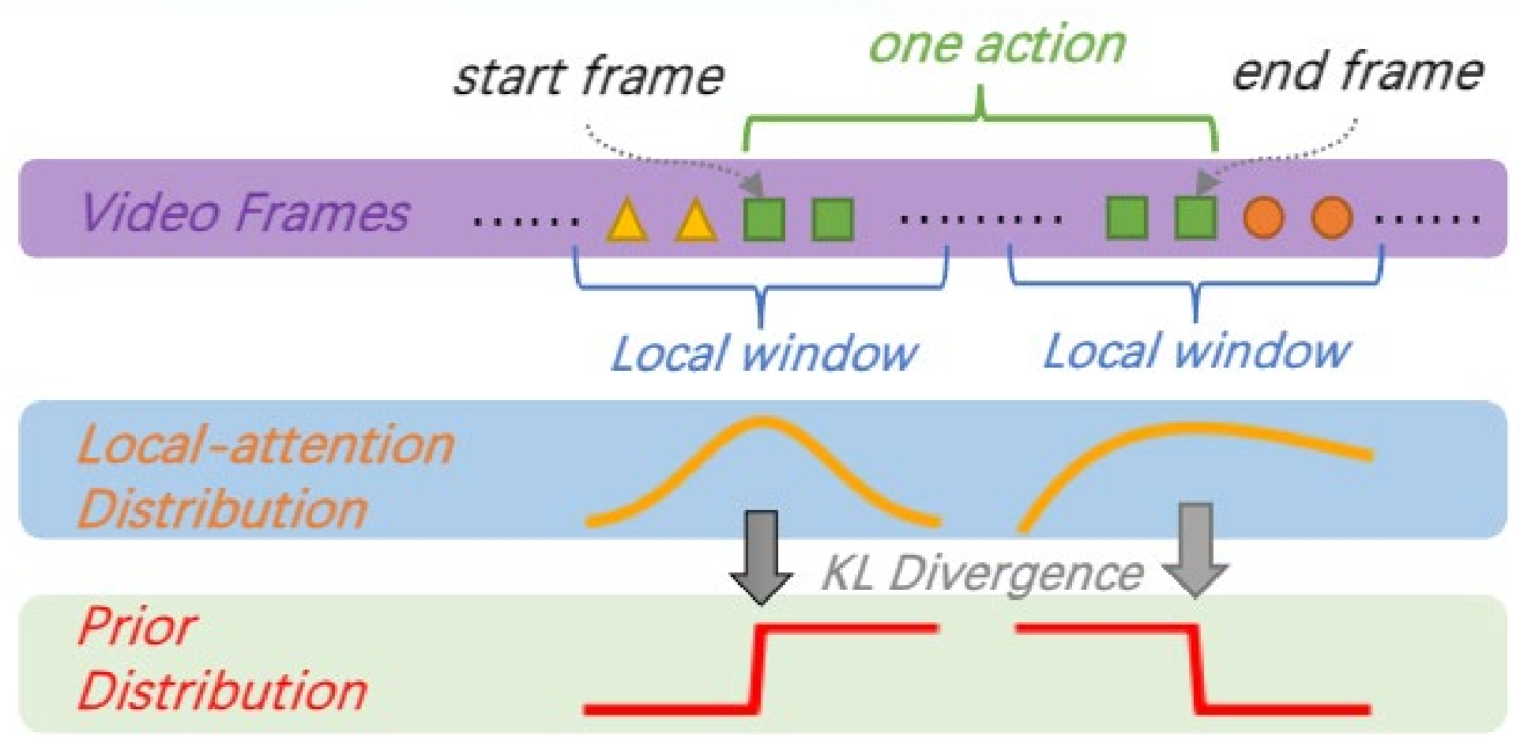
2. 研究内容与创新

本技术在动作分割领域首次提出了一个没有时间卷积模块的纯Transformer模型，称为TUT，如下图所示。该模型将局部注意力和时间上下采样融合到Transformer中，形成一个类U-Net的架构。因此，模型不仅降低了复杂度，还摆脱了不灵活的时间卷积模块对性能的限制。



TUT模型架构

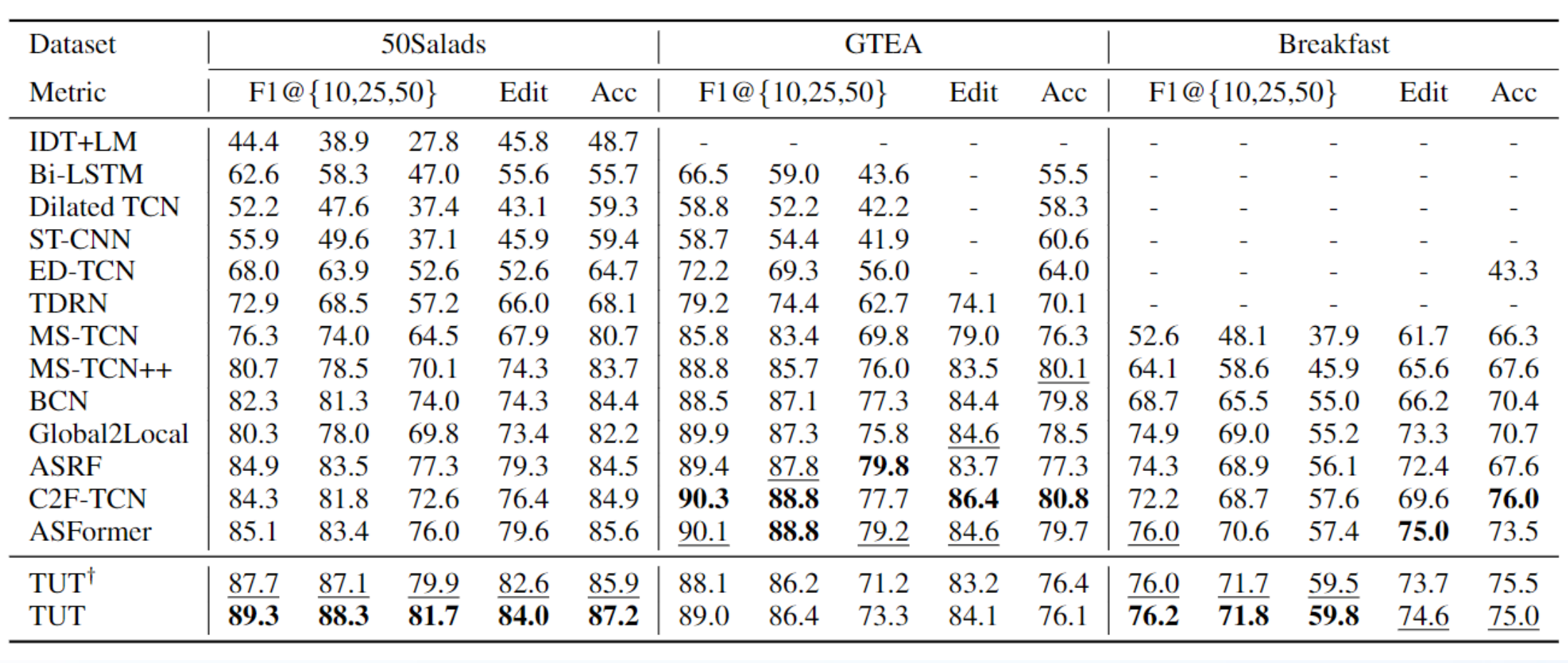
此外，基于来自注意模块和边界标签的帧间相似度得分分布，我们提出了一种基于分布的边界感知损失，使我们的模型能够更准确地分类边界。



边界感知损失

3. 实验与结论

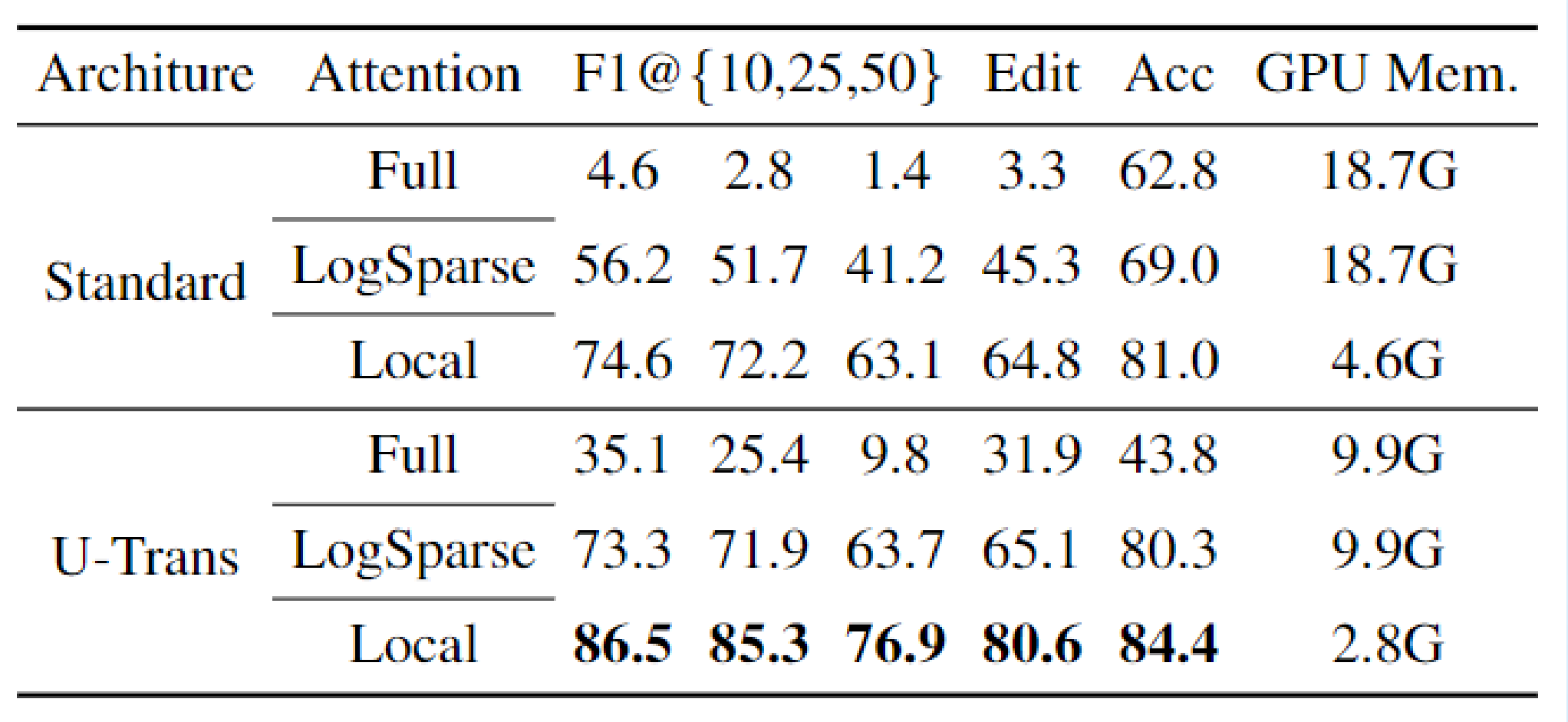
（1）与之前的深度学习方法进行对比



Dataset	50Salads					GTEA					Breakfast				
Metric	F1@{10,25,50}			Edit	Acc	F1@{10,25,50}			Edit	Acc	F1@{10,25,50}			Edit	Acc
IDT+LM	44.4	38.9	27.8	45.8	48.7	-	-	-	-	-	-	-	-	-	-
Bi-LSTM	62.6	58.3	47.0	55.6	55.7	66.5	59.0	43.6	-	55.5	-	-	-	-	-
Dilated TCN	52.2	47.6	37.4	43.1	59.3	58.8	52.2	42.2	-	58.3	-	-	-	-	-
ST-CNN	55.9	49.6	37.1	45.9	59.4	58.7	54.4	41.9	-	60.6	-	-	-	-	-
ED-TCN	68.0	63.9	52.6	52.6	64.7	72.2	69.3	56.0	-	64.0	-	-	-	-	43.3
TDRN	72.9	68.5	57.2	66.0	68.1	79.2	74.4	62.7	74.1	70.1	-	-	-	-	-
MS-TCN	76.3	74.0	64.5	67.9	80.7	85.8	83.4	69.8	79.0	76.3	52.6	48.1	37.9	61.7	66.3
MS-TCN++	80.7	78.5	70.1	74.3	83.7	88.8	85.7	76.0	83.5	80.1	64.1	58.6	45.9	65.6	67.6
BCN	82.3	81.3	74.0	74.3	84.4	88.5	87.1	77.3	84.4	79.8	68.7	65.5	55.0	66.2	70.4
Global2Local	80.3	78.0	69.8	73.4	82.2	89.9	87.3	75.8	84.6	78.5	74.9	69.0	55.2	73.3	70.7
ASRF	84.9	83.5	77.3	79.3	84.5	89.4	87.8	**79.8**	83.7	77.3	74.3	68.9	56.1	72.4	67.6
C2F-TCN	84.3	81.8	72.6	76.4	84.9	**90.3**	**88.8**	77.7	**86.4**	**80.8**	72.2	68.7	57.6	69.6	**76.0**
ASFormer	85.1	83.4	76.0	79.6	85.6	90.1	**88.8**	79.2	84.6	79.7	76.0	70.6	57.4	**75.0**	73.5
TUT†	87.7	87.1	79.9	82.6	85.9	88.1	86.2	71.2	83.2	76.4	76.0	71.7	59.5	73.7	75.5
TUT	**89.3**	**88.3**	**81.7**	**84.0**	**87.2**	89.0	86.4	73.3	84.1	76.1	**76.2**	**71.8**	**59.8**	74.6	75.0

结论：在两个比较大的数据集上，TUT模型几乎在所有指标上都超过了之前的模型。而且，在边界感知损失的共同训练下，模型的性能可以进一步提升，这证明了边界感知损失的有效性。在最小的数据集上，TUT模型的性能略落后于之前最好的模型。这是合理的，因为小数据集难以充分训练纯Transformer模型

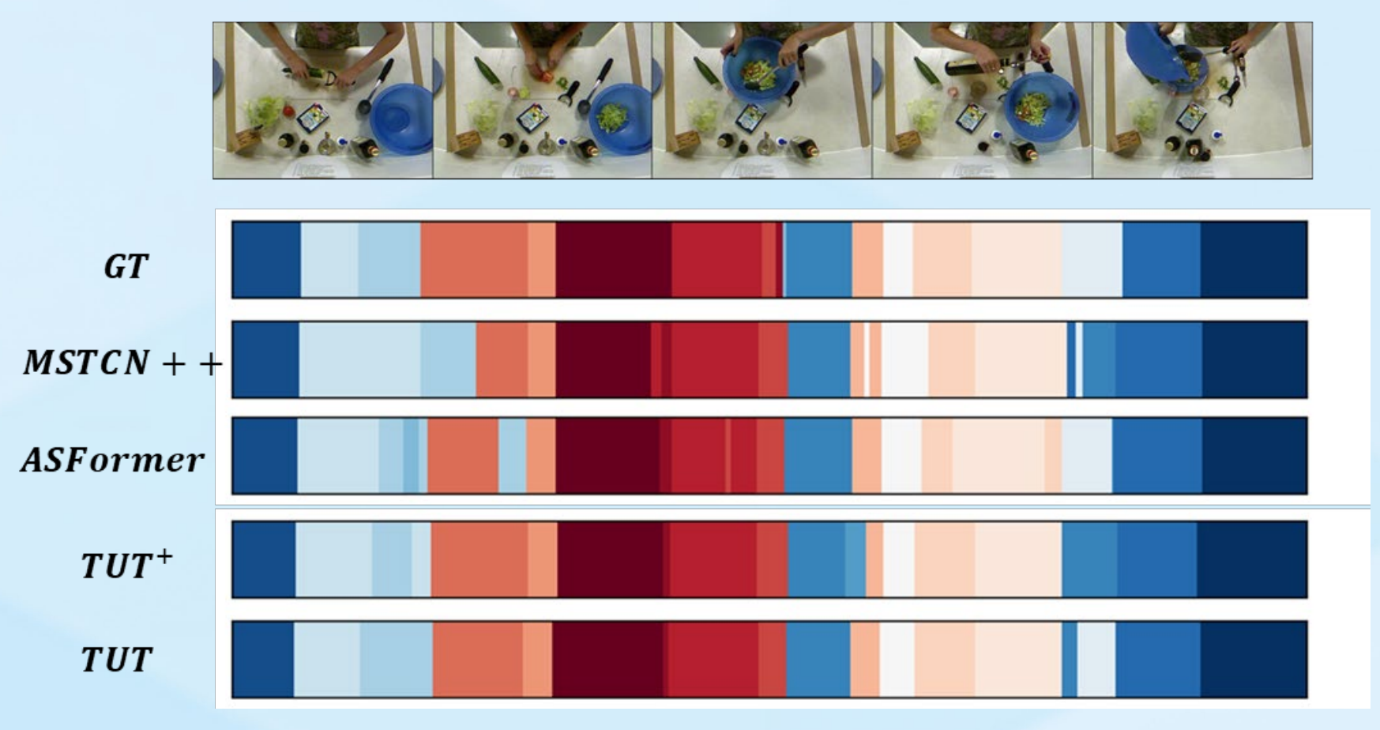
（2）模型结构和注意力的消融实验



Architure	Attention	F1@{10,25,50}			Edit	Acc	GPU Mem.
Standard	Full	4.6	2.8	1.4	3.3	62.8	18.7G
	LogSparse	56.2	51.7	41.2	45.3	69.0	18.7G
	Local	74.6	72.2	63.1	64.8	81.0	4.6G
U-Trans	Full	35.1	25.4	9.8	31.9	43.8	9.9G
	LogSparse	73.3	71.9	63.7	65.1	80.3	9.9G
	Local	**86.5**	**85.3**	**76.9**	**80.6**	**84.4**	2.8G

结论：带有时间采样的Transformer架构的性能优于标准Transformer架构，同时GPU内存消耗更少。不管架构如何，完全注意力的效果都很差，这表明对小数据的训练需要更稀疏的注意力模式。由于在动作分割中相邻帧通常具有较强的相关性，所以采用局部注意力机制的要比对数稀疏注意力机制好得多。

（3）分割效果展示



50Salads数据集上分割效果

结论：根据多种模型的分割效果对比可以验证TUT模型的优越性，同时带有边界感知训练的TUT产生了更好的性能，证明了边界感知的有效性。