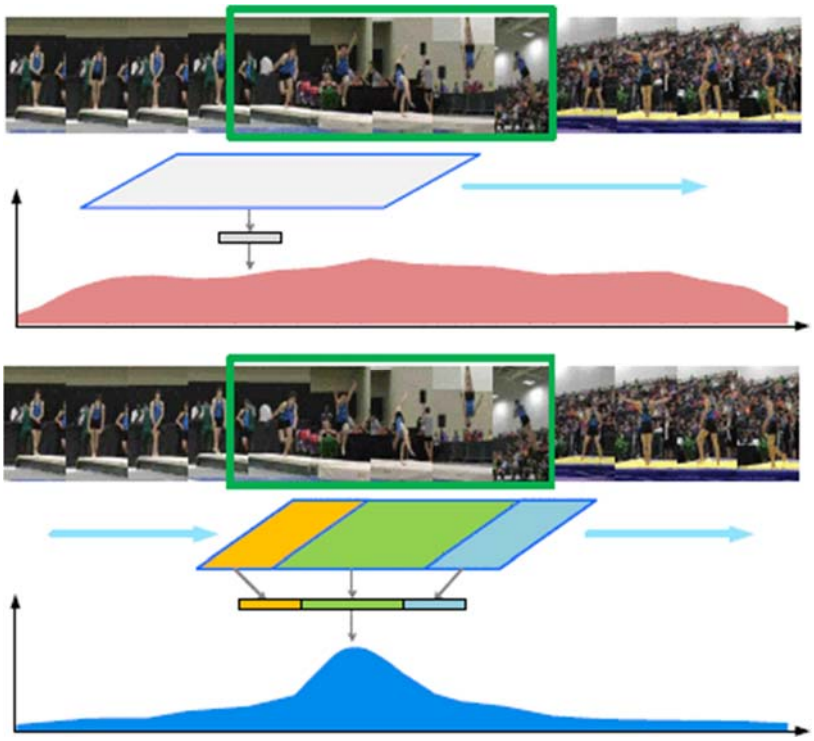
2018 年秋-第 10 周

Temporal Action Detection with Structured Segment Networks

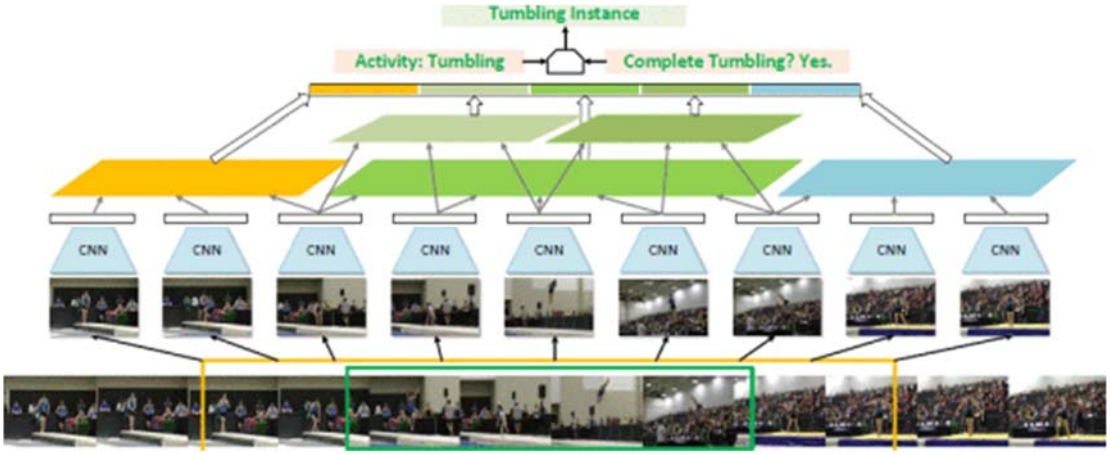
1、SSN 的优势



上图：探测器构建的功能没有任何阶段结构的动作，例如整个窗口的平均合并。每当它看到任何与翻滚相关的判别性片段时，它就会产生高响应，这使得很难本地化实例。

底部：SSN 探测器通过结构化时间金字塔池化利用阶段结构（起始，过程和结束）。当窗口对齐良好时，它的响应才显着。

2、SSN 结构概述



来自 ActivityNet 的视频有一个候选区域（绿框）。我们首先通过扩展它来构建扩充提议（黄色框）。增强的 proposal 分为 starting（橙色），course（绿色）和 ending（蓝色）阶段。在 course 阶段构建了具有两个子部分的附加级别的金字塔。来自 cnns 的

特征汇集在这五个部分中并连接起来形成全局区域表示。活动分类器和完整性分类器对区域表示进行操作以产生活动概率和类条件完整性概率。proposal 为正实例的最终概率由这两个分类器的联合概率决定。在训练期间，我们稀疏地采样 L=9 来自均匀分割的片段的片段以近似密集的时间金字塔池。

3、SSN 流程

从输入到输出，需要三个关键步骤。

首先,框架依赖于 proposal 方法来生成一组具有不同持续时间的 temporal proposals，其中每个提议都带有开始和结束时间。我们的框架将每个 proposal 视为三个连续阶段的组合，即 starting，course，ending，分别捕捉动作如何开始，继续和结束。

因此，对于每个提议，结构化时间金字塔池（STPP）通过

1）将提议分成三个阶段来执行;

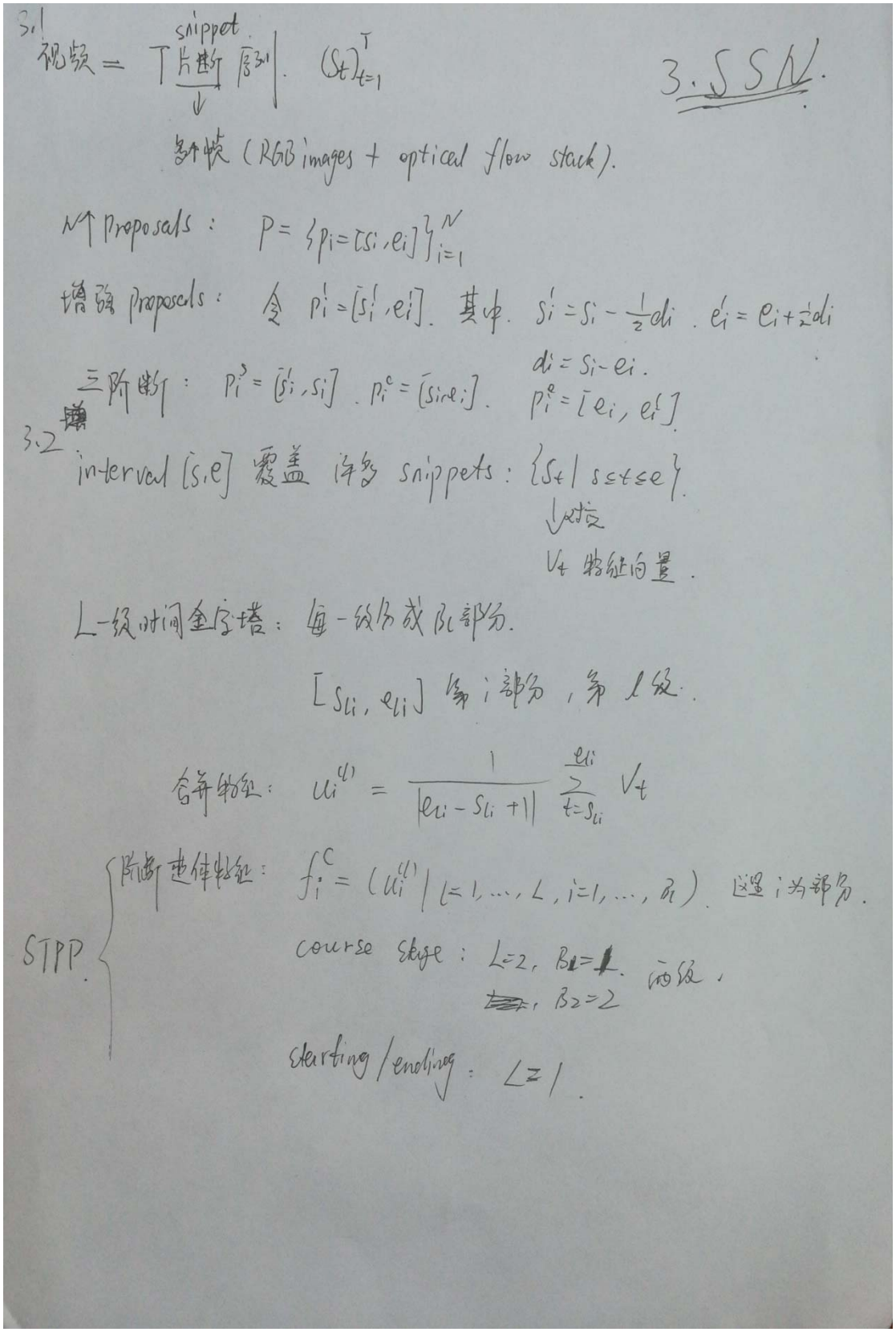
2）为每个阶段建立时间金字塔表示;

3）通过连接阶段级表示来构建整个提案的全局表示。

最后,分别用于识别活动类别和评估完整性的两个分类器将应用于由 STPP 获得的表示，并且它们的预测将被组合，从而产生用类别标签标记的完整实例的子集。其他提案，被视为属于背景或不完整，将被过滤掉。

上面列出的所有组件都集成到一个统一的网络中，该网络将以端到端的方式进行培训。对于训练，我们采用稀疏片段采样策略来近似密集样本上的时间金字塔。通过利用视频片段之间的冗余，该策略可以显着降低计算成本，从而允许对长期时间结构进行关键建模。

4、SSN 结构详解


3.1
视频 = T 片断 snippet 序列 $(S_t)_{t=1}^{T}$
3. SSN.
多帧 (RGB images + optical flow stack).
N个 proposals: $P=\{p_i=[s_i,e_i]\}_{i=1}^{N}$
增强 proposals: 令 $p_i'=[s_i',e_i']$. 其中 $s_i'=s_i-\frac{1}{2}d_i$. $e_i'=e_i+\frac{1}{2}d_i$
$d_i=s_i-e_i$.
三阶段: $p_i^s=[s_i',s_i]$. $p_i^c=[s_i,e_i]$. $p_i^e=[e_i,e_i']$
3.2
interval $[s,e]$ 覆盖 许多 snippets: $\{s_t \mid s\le t\le e\}$
对应
V_t 特征向量.
L-级时间金字塔: 每一级分成 B_l 部分.
$[s_{li},e_{li}]$ 第 i 部分, 第 l 级.
合并特征: $u_i^{(l)}=\frac{1}{|e_{li}-s_{li}+1|}\sum_{t=s_{li}}^{e_{li}} V_t$
STPP
阶段 连体特征: $f_i^c=(u_i^{(l)} \mid l=1,\dots,L, i=1,\dots,B_l)$. 这里 i 为部分.
course stage: $L=2$, $B_1=1$. 两级.
$B_2=2$
starting / ending: $L=1$.

3.3.

活动分类器：input proposals ⟶ k+1 类. { K 活动类 (1→K)
                                          1 背景类 (0).

阻断在 ~~feature~~ course 阶段. 利用 f_i^c.

完整性分类器：$\{C_k\}_{k=1}^{K}$ ⟶ 对应 K 个活动类.

基于 $\{f_i^s, f_i^c, f_i^o\}$ 判断是否完整.

统一分类损失（条件概率推导）：

$$L_{cls}(c_i, b_i; p_i) = -\log P(c_i|p_i) - I(c_i>1)\log P(b_i|c_i, p_i)$$

其中：c_i 类标签. b_i 表示是否完整.

训练样本：(1). 正样本：与 GT 的 IOU ≥ 0.7 → $c_i>0$, $b_i=1$

(2) 背景：与任何 GT 不重叠 → $c_i=0$

(3) 不完整：自身跨在 GT 中，但 IOU < 0.3 → $c_i>0, b_i=$

3.4 位置回归 $\{R_k\}_{k=1}^{K}$ ⟶ 对应 K 个活动类.

对 p_i：回归相对变化：中心 μ_i

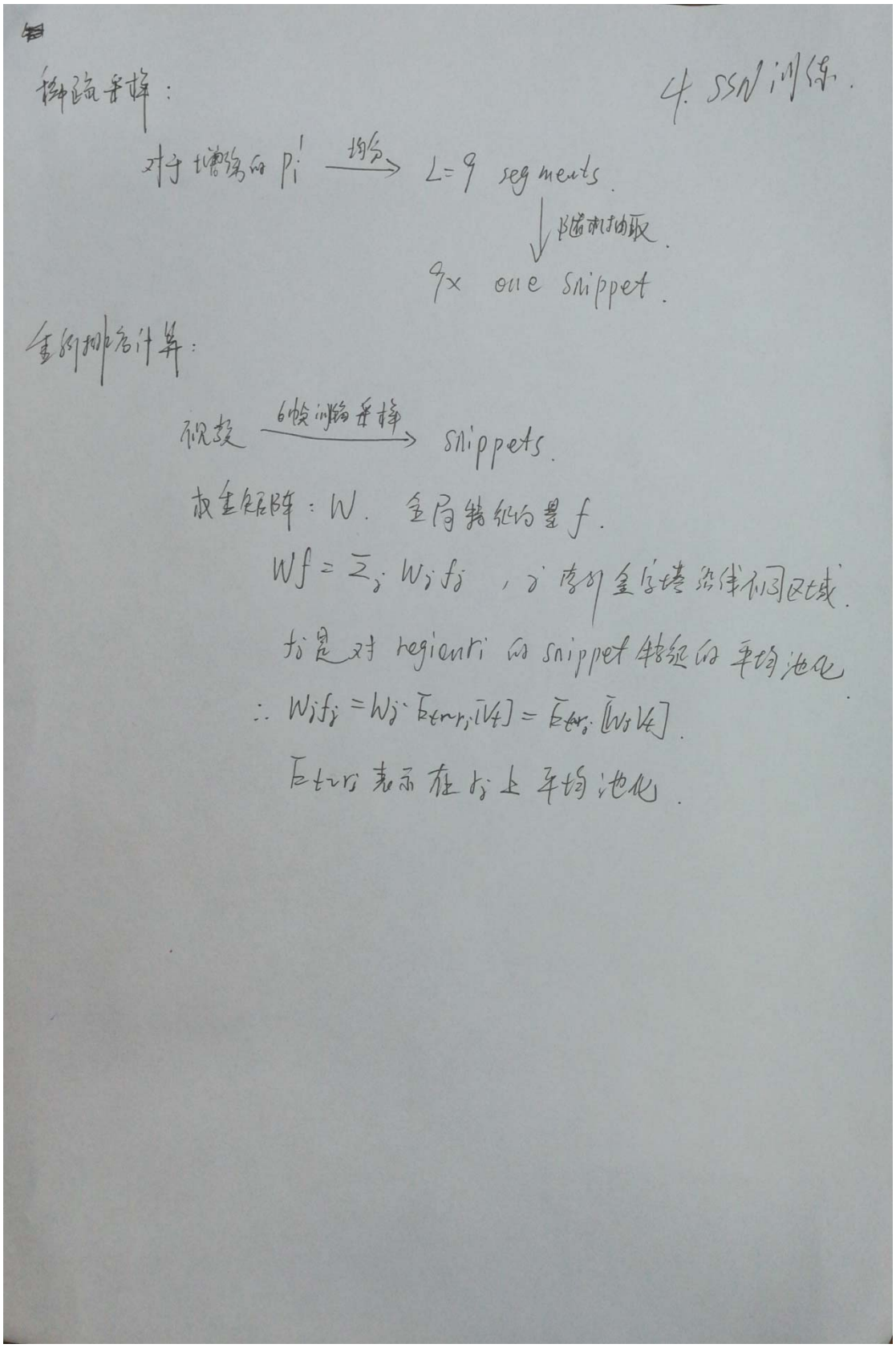
跨度 ϕ_i（log 尺度）.

使用最近 GT 为回归

多任务损失：

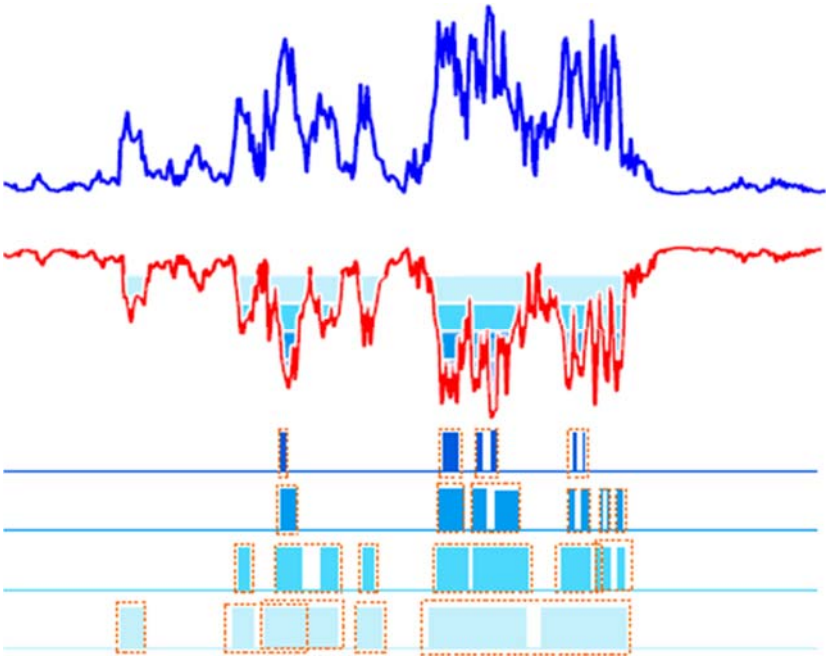
$$L_{cls}(c_i, b_i; p_i) + \lambda \cdot I(c_i>1 \,\&\, b_i=1) L_{reg}(\mu_i, \phi_i; p_i)$$

5、SSN 训练


稀疏采样：
4. SSN 训练.
对于增强后的 P_i' → 均分 → $L=9$ segments.
↓ 随机抽取.
9× one snippet.
金字塔各计算：
视频 → 6帧间隔采样 → snippets.
权重矩阵：W. 全局特征向量 f.
$Wf = \sum_j W_j f_j$，j 为金字塔各层不同区域.
于是对 region i 的 snippet 特征的平均池化.
$\therefore W_i f_i = W_i \cdot E_{str,i}[V_i] = E_{str,i}[W_i V_i]$.
Estr,i 表示在 i 上平均池化.

6、proposals 的生成（TAG）

分水岭算法思想



可视化提案生成的时间动作性分组过程。

上：作为一维信号序列的动作概率。

中：补充信号。我们用不同的水平 γ 淹没它。

底部：通过不同的洪水水平获得的区域。

通过根据分组标准合并区域，我们得到最终的 proposal 集（橙色）。

算法在不同“ 水位 ”（γ）的地形上淹水，导致一组被水覆盖的“盆地”，表示为 G（γ）。

分组方案的工作原理：

从一个种子盆地开始，并连续吸收随后的盆地，直到盆地持续时间超过总持续时间（即从第一个盆地开始到最后一个盆地结束）的部分下降一定的门槛 τ。这句话直观来说就是水填满了后在满了的基础上下降 τ。

然后将吸收的盆地和它们之间的空白区域分组以形成单个 proposal。

将每个盆地视为种子并执行分组程序以获得一组表示的提议 G′（τ，γ）。用步长为 0.05 统一采样 τ， γ ∈（0，1）。这两个阈值的组合导致多组 regions。然后，我们联合他们。

最后，我们将非最大抑制应用于具有 IoU 阈值 0.95 的并集，以过滤掉高度重叠的 proposal。保留的 proposal 将提供给 SSN 框架。

7、实验

ActivityNet ：两个版本 v1.2 和 v1.3.前者包含 9682 个视频，100 个分类。后者包含 19994 个视频，200 个分类。分成三部分：training， validation， testing，比例 2:1:1

THUMOS14：1010 个 validation 视频，1574 个 testing 视频。20 个分类。Training set 为 UCF101。一般将 validation 作为训练集，因此，训练集包含 220 个视频，20 个分类。

Implementation Details：在每一个 minibatch 中，positive background incomplete proposals 的比例为 1:1:6。