Siamese RPN

- 特点
  - 用大量的 image pairs end to end 训练
    - 并不需要连续的视频流
  - Inference 阶段相当于执行一个 one-shot 检测任务
- 网络结构
  - Feature extraction 子网络
    - template branch, detection branch
    - 两个分支共享 CNN 参数
  - Region proposal 子网络
    - classification branch, regression branch

Siamese RPN

- 网络结构
  - template feature map 当作卷积核来使用

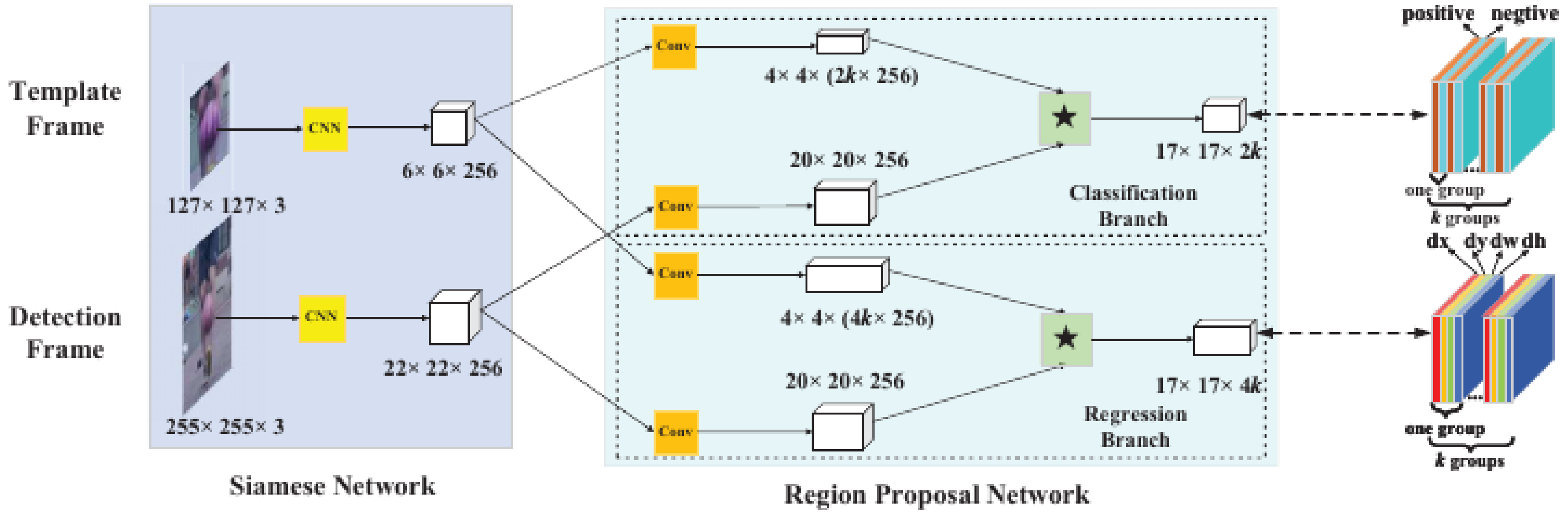
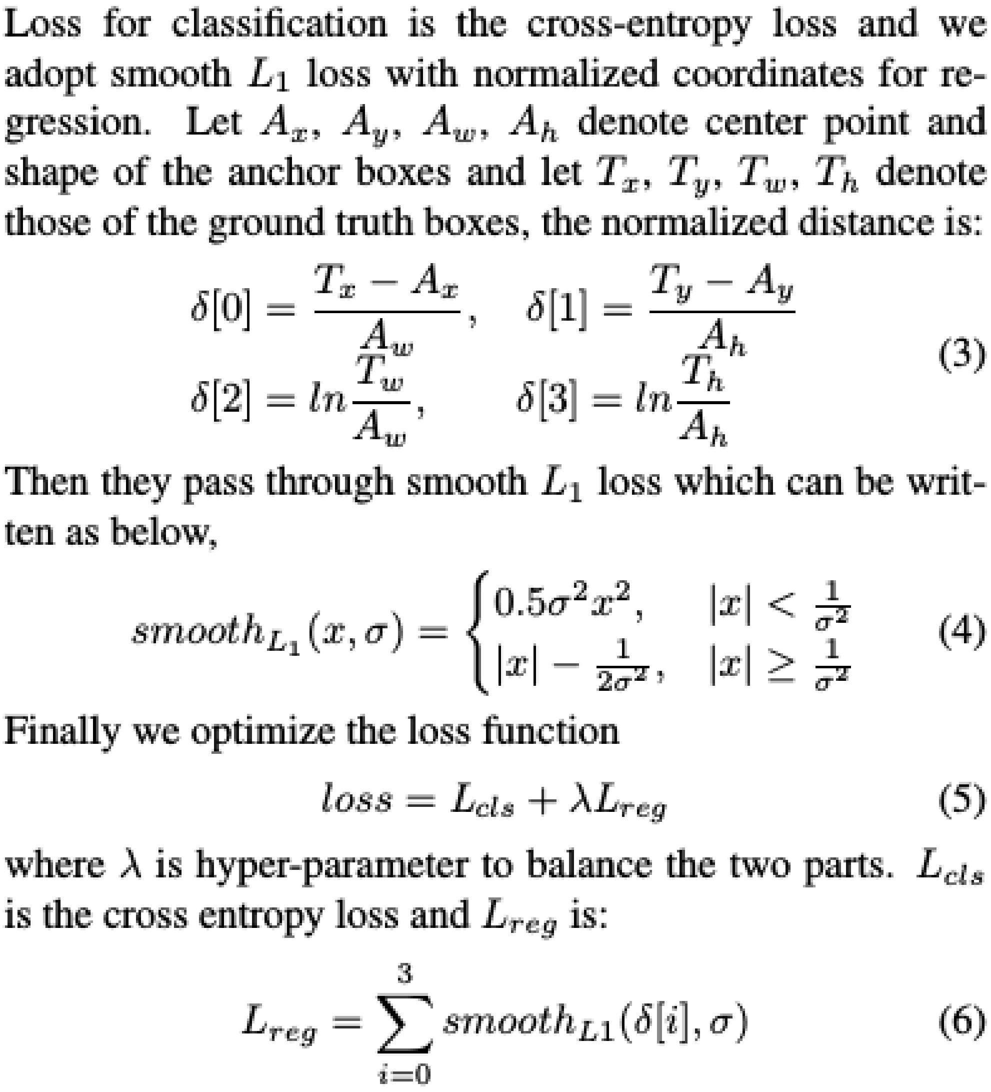


Figure 2: Main framework of Siamese-RPN: left side is Siamese subnetwork for feature extraction. Region proposal subnetwork lies in the middle, which has two branches, one for classification and the other for regression. Pair-wise correlation is adopted to obtain the output of two branches. Details of these two output feature maps are in the right side. In classification branch, the output feature map has $2k$ channels which corresponding to foreground and background of k anchors. In regression branch, the output feature map has $4k$ channels which corresponding to four coordinates used for proposal refinement of k anchors. In the figure, $\star$ denotes correlation operator.

Siamese RPN

- Loss
  - 分类用交叉熵
  - 回归用 smooth L1
- Training
  - positive samples
    - anchors
    - IoU > 0.6
  - negative samples
    - anchors
    - IoU < 0.3



Loss for classification is the cross-entropy loss and we adopt smooth L_1 loss with normalized coordinates for regression. Let A_x, A_y, A_w, A_h denote center point and shape of the anchor boxes and let T_x, T_y, T_w, T_h denote those of the ground truth boxes, the normalized distance is:

$$\begin{aligned} \delta[0] &= \frac{T_x - A_x}{A_w}, \quad \delta[1] = \frac{T_y - A_y}{A_h} \\ \delta[2] &= ln\frac{T_w}{A_w}, \quad \delta[3] = ln\frac{T_h}{A_h} \end{aligned} \tag{3}$$

Then they pass through smooth L_1 loss which can be written as below,

$$smooth_{L_1}(x, \sigma) = \begin{cases} 0.5\sigma^2 x^2, & |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2}, & |x| \geq \frac{1}{\sigma^2} \end{cases} \tag{4}$$

Finally we optimize the loss function

$$loss = L_{cls} + \lambda L_{reg} \tag{5}$$

where λ is hyper-parameter to balance the two parts. L_{cls} is the cross entropy loss and L_{reg} is:

$$L_{reg} = \sum_{i=0}^{3} smooth_{L1}(\delta[i], \sigma) \tag{6}$$

Siamese RPN

- Inference
    - 在初始帧计算 template 的卷积核，并且在整个跟踪过程中保持不变
    - 用当前帧的 feature map 和 template 的卷积核进行卷积，得到 classification 和 regression 输出
    - 生成 top K proposals
    - Proposal selection
        - 先丢弃距离 feature map 中心太远的 bounding boxes
            - 基于相邻帧运动距离不会太远的假设
        - 再使用惩罚项重排 proposals，取最优的结果

Siamese RPN

- Inference

- 计算 bbox

$$x_i^{pro} = x_i^{an} + dx_l^{reg} * w_l^{an}$$
$$y_j^{pro} = y_j^{an} + dy_l^{reg} * h_l^{an}$$
$$w_l^{pro} = w_l^{an} * e^{dw_l}$$
$$h_l^{pro} = h_l^{an} * e^{dh_l}$$

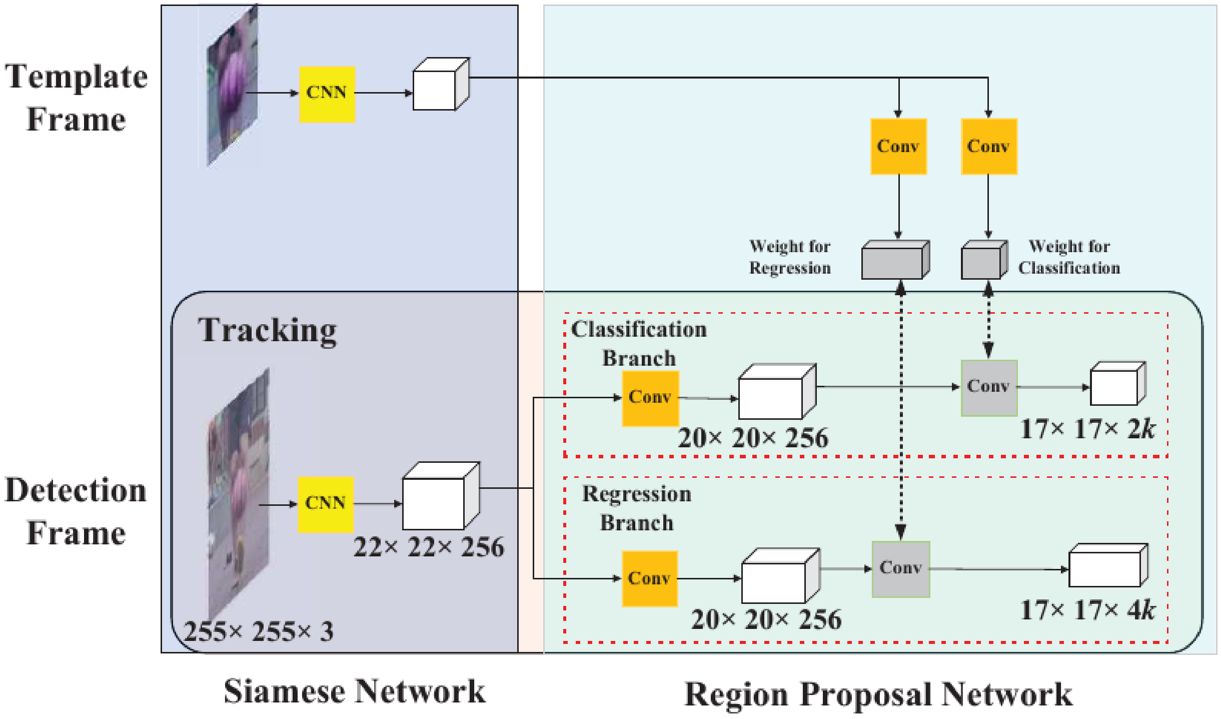


Figure 3: Tracking as one-shot detection: the template branch predicts the weights(in gray) for kernels of region proposal subnetwork on detection branch using the first frame. Then the template branch is pruned and only the detection branch is retained. So the framework is modified to a local detection network.

Siamese RPN

- 实验
  - 测试数据集
    - VOT2015 (60 videos)
    - VOT2016 (60 videos)
    - VOT2017 real-time (60 videos)
    - OTB2015 (100 videos)
  - 测试指标
    - EAO (Expected Average Overlap)
      - both accuracy and robustness
    - Precision and success plot
  - 测试结果
    - 又快又准， state of the art