

硕 士 研 究 生 读 书 报 告



题目 High Performance Visual Tracking with Siamese Region Proposal Network

作者姓名 刘尚楠

作者学号 21851164

指导教师 李启雷

学科专业 软件工程

所在学院 软件学院

提交日期 二○一八年十二月

摘要

论文认为目前存在的跟踪模型很难在保证实时性的前提下达到最好的最优的性能，因此该论文提出了Siamese-RPN网络，该网络能够利用大规模的数据集进行端到端的离线训练。Siamese-RPN网络包含（1）特征提取（2）区域建议网络两个部分。同时在推理阶段，利用meta-learning思想将跟踪任务作为 local one-shot detection task（具体细节见下文）。由于该论文利用了目标检测过程中的RPN网络对跟踪位置进行了调优，摒弃了常见的基于Siamese-FC模型进行多尺度的训练部分，因此不但达到了160FPS的速度，同时在VOT2015, VOT2016 and VOT2017实现了好的性能。

**关键词**：RPN，图像跟踪

Abstract

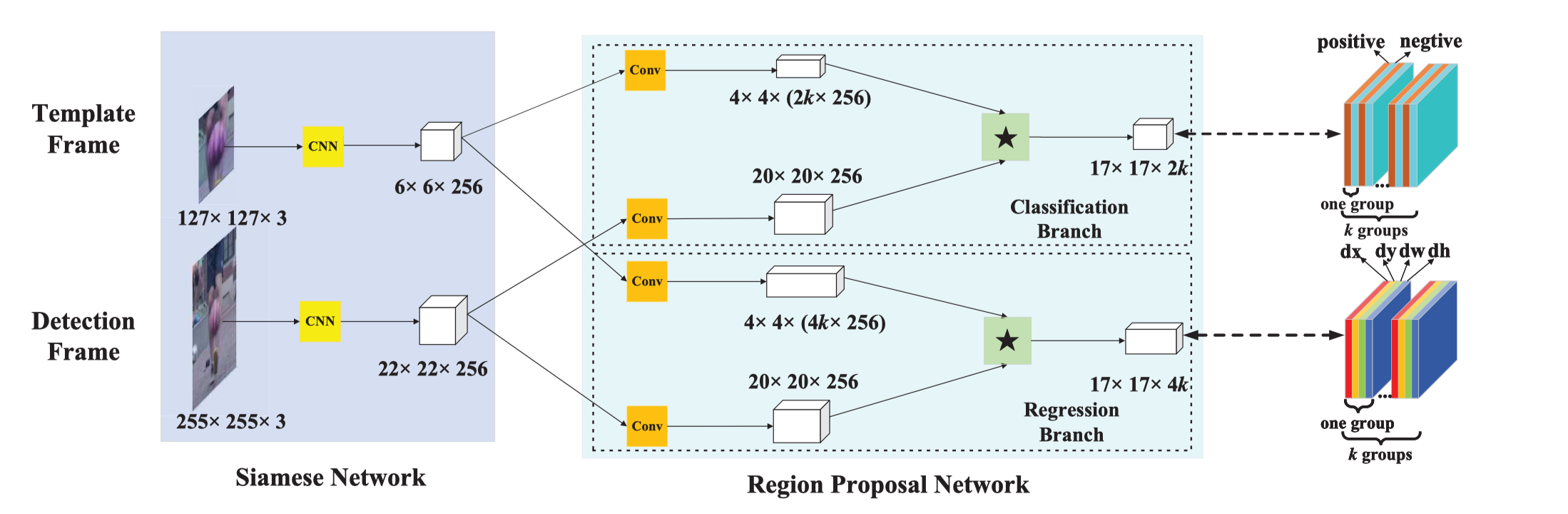
Visual object tracking has been a fundamental topic in recent years and many deep learning based trackers have achieved state-of-the-art performance on multiple benchmarks. However, most of these trackers can hardly get top performance with real-time speed. In this paper, we propose the Siamese region proposal network (Siamese-RPN) which is end-to-end trained off-line with large-scale image pairs. Speciﬁcally, it consists of Siamese subnetwork for feature extraction and region proposal subnetwork including the classiﬁcation branch and regression branch. In the inference phase, the proposed framework is formulated as a local one-shot detection task. We can pre-compute the template branch of the Siamese subnetwork and formulate the correlation layers as trivial convolution layers to perform online tracking. Beneﬁt from the proposal reﬁnement, traditional multi-scale test and online ﬁne-tuning can be discarded. The Siamese-RPN runs at 160 FPS while achieving leading performance in VOT2015, VOT2016 and VOT2017 real-time challenges.

**Keywords**：RPN, Target tracking

1 动机

自KCF以后，基于相关滤波的目标跟踪方法是近几年研究的热点。而从fasterrcnn、Maskrcnn以来，目标检测框架基本也就定型。然而目标跟踪和检测具有很大的性，既然基于深度学习的目标检测已经发展的相对成熟，因此作者认为通过合理的设计网络，将检测的思想迁移至跟踪上也应该会不错的性能。

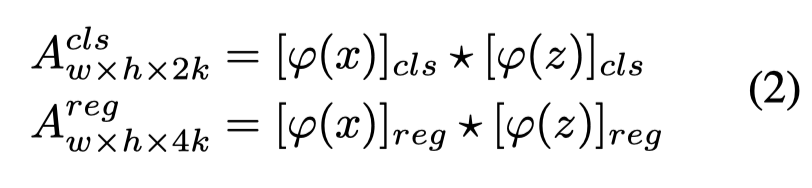
2 训练



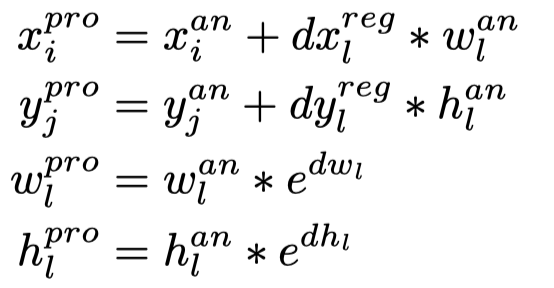
框架主要分为两部分：

1. 用于目标特征提取的Siamese Network
2. 位置回归的RPN网络。

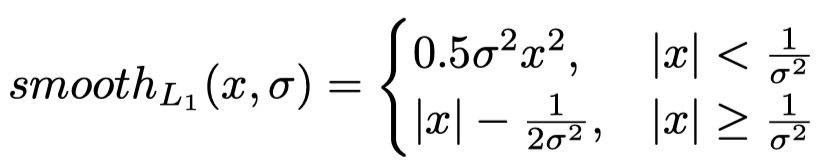
第一部分和SiameseFC完全一样。因此框架的精髓包含在了第二部分。模板帧在提取特征之后被一个3\*3的卷积核（推测）分成两个分支，假设有k个anchors，则第一个分支用于分类（前景和背景），输出维度为4\*4\*(2k\*256)，第二分支用于位置回归，输出维度为4\*4(4k\*256)；检测帧经过（1）后经过3\*3\*256的卷积核后，输出20\*20\*256的两个分支和 。最后两个部分分别进行分组卷积操作：

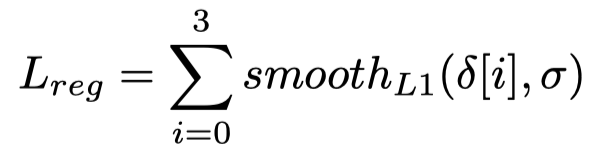


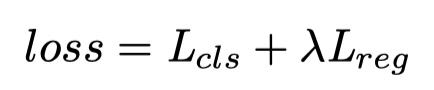
分别代表前景和背景的概率, 表示与真实目标框的测量距离（dx, dy, dw, dh）。论文选用IOU>0.6的为正样本，IOU<0.3的为负样本，在每一个batch(64)中控制正样本最多为16个。则正样本排序后选择前K个预测框作为输出：



表示上的位置在模板帧上的对应位置。训练过程的分类损失为交叉熵损失函数，回归损失函数为。

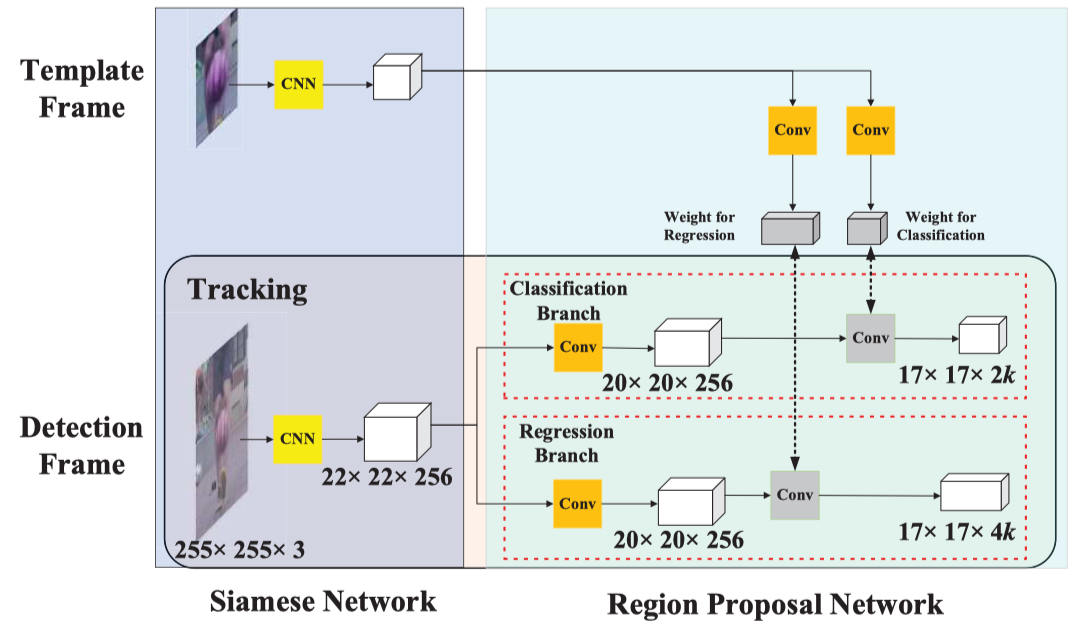






关键点

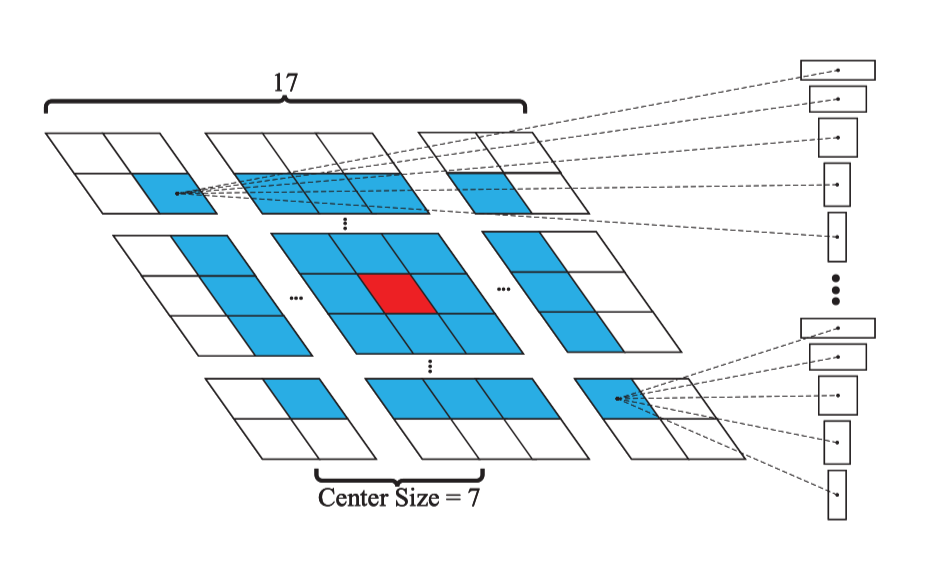
跟踪阶段借鉴meta-learning的思想，将跟踪任务作为local one-shot detection任务，从而避免了在线更新（在之后的版本中好像加了在线跟新的策略，在不牺牲速度的前提下，提高了性能，这样才可能适用于long time 序列）。



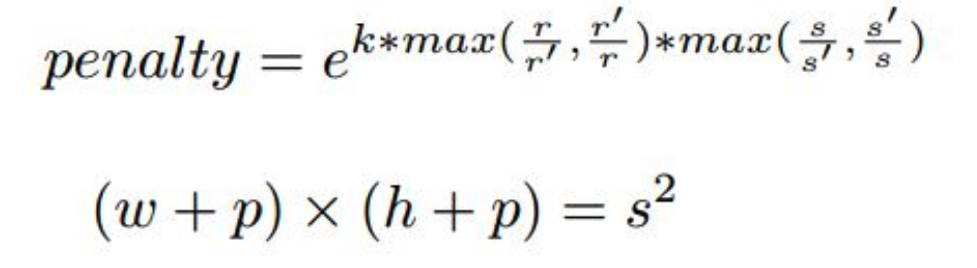
通俗的来讲，就是在训练完成后，只将第一帧预测的模板（分类和回归）作为检测帧的相关核，以致该模型具有较高的速度。

测试过程中的trick

代替使用w \* h\*k个anchors,作者在实验的过程中使用半径为g的g\*g\*k个anchors。因为相邻帧之间位移通常较小，这样可以舍弃一些异常点。



在（1）的基础上对加了余弦窗和尺度（scale）惩罚项。余弦窗抑制大位移的偏移，惩罚项可以抑制anchor的大小(size)和宽高比（ratio）



P = (w+h)/2

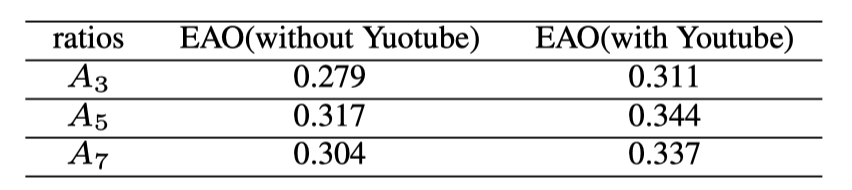
r和分别表示建议的和当前的bounding box高宽比，s和同理表示scale，k为超参数。

经过(1)和(2)之后我们对重新排序，通过NMS后得到目标框的位置。

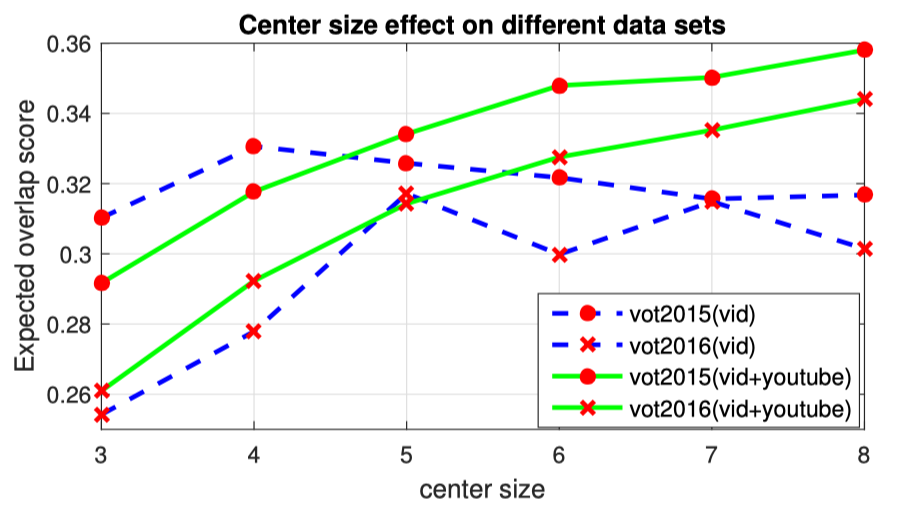
实验

数据来源：实验过程中我们抽取图像对从ILSVRC和Youtube-BB数据集，间隔小于100帧。

训练过程中的anchor：与目标检测中标准RPN过程不同，跟踪模型中的scale固定（文章未给出大小），ratios作者试了三种[0.5, 1, 2], [0.33, 0.5, 1, 2, 3],[0.25, 0.33, 0.5, 1, 2, 3, 4]（A3, A5, A7），通过实验发现A5最好。



上文当中g的选择作者也做了实验，证明g的大小与数据集的大小有关。



OTB2015、VOT2016、VOT2017 论文做了详细的实验，可以看文章。

论文贡献

（1）可以在大规模的跟踪数据集上进行end-to-end的离线训练

（2）摒弃了多尺度前向卷积过程

（3）在跟踪模型中使用one-shot learning 思想

（4）将RPN网络与SiameseFC很好的结合，达到了好的性能。

参考文献

[1] High Performance Visual Tracking with Siamese Region Proposal Network Bo Li, Junjie Yan,Wei Wu1 , Zheng Zhu1, Xiaolin Hu. CVPR, 2018.