**Online Track**

VOS：Vitual Object Segmentation

VOT: Vitual Object Tracking

IOU: Jaccard index

参考链接：<https://zhuanlan.zhihu.com/p/58154634>

代码地址：<https://github.com/foolwood/SiamMask>

mAP（mean of Average Precision）：<https://zhuanlan.zhihu.com/p/56961620>

[Online Track](#8221-1586187781202)

[难点](#6440-1587483456963)

[SiamMask简介](#8790-1586240945566)

[三个任务](#5339-1586241816947)

[训练简介](#6210-1586245782781)

[关联工作](#6947-1586259262819)

[两个问题](#5243-1586259289903)

[方法](#5747-1586790472978)

[参考的全卷积Siamese框架](#4244-1586790564196)

[SiamMask](#3798-1586842730282)

[补充细节](#2370-1587483218213)

[实验](#2527-1587621865343)

[在vot-2016和vot-2018上评估](#5534-1587633040683)

[对象表示的重要性](#3276-1587633408462)

[在VOT-2018和VOT-2016上的结果](#2524-1587892196091)

[在半监督VOS上的评估](#1430-1587896171808)

[在DAVIS和YouTube-VOS上的结果](#4699-1587908861567)

[深入分析](#3047-1587912049972)

[结构细节](#4218-1587913195464)

[网络主干](#9543-1587914886623)

[网络头](#8351-1587916008765)

[掩码细化模块](#3651-1587916102892)

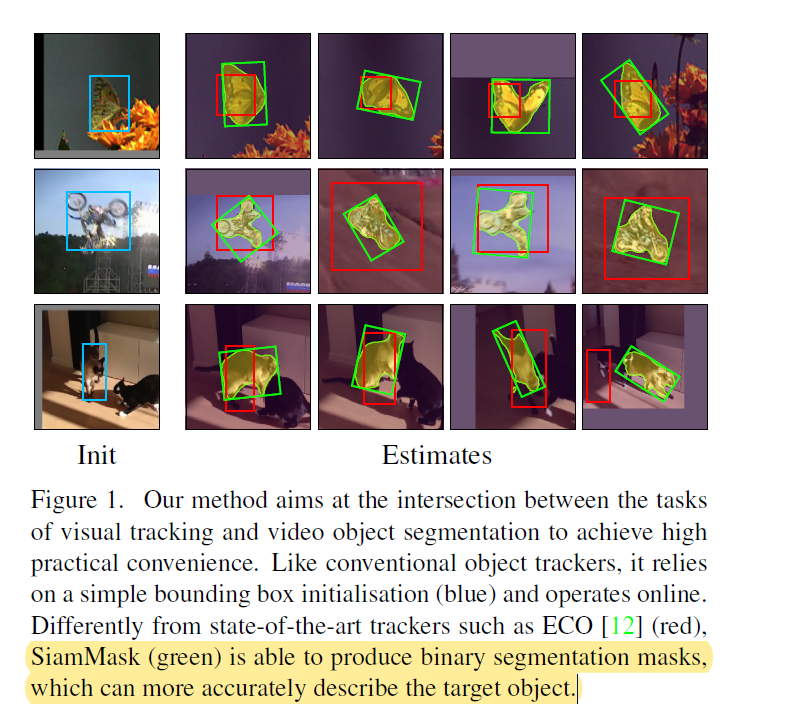
[进一步定性结果](#3024-1587916610072)

**难点**

可在线跟踪，不能使用视频未来的帧来判断目标现在的位置

视频的第一帧很重要，类似地，VOS（视频分割）要求估计视频第一帧任意目标的位置，但需要很多的计算时间。

最快的方法也无法实时计算



**SiamMask简介**

多任务学习，学习锚框和图像分割。

基于全卷积的Siamese网络，并在线下学习大量的视频（VOS）

目标是在保持线下训练能力和在线学习速度的同时，显著改变他们对目标图像的表示

**三个任务**

在流动窗口学习目标对象和多个候选对象之间的相似性度量，输出是一个密集响应图，该图只表面目标的位置，不提供空间幅度的信息

使用网络区域候选网络来学习锚框回归

类别不可知二元分割，二元标签只在线下计算分割损失的时候被需要

**以上每个任务被每个不同的分支所代表，分支从一个共享CNN上分离，最后贡献给一个loss，它汇总三个输出。**

**训练简介**

仅仅依赖一个单锚框初始化

不需要更新进行在线计算，并生成对象分割掩码和每秒55帧的旋转锚框

没有在微调、数据增强、光流等昂贵技术花费

**关联工作**

**两个问题**

**视觉对象跟踪**。

过去的相关滤波器算法允许区分任意目标的模板和它的二维翻译。新方法在数对视频帧上训练一个在线的相似函数，在测试的时候，这个函数可以评估一个新的视频，一次一帧。特别是全卷积Siamese方法的改进，利用区域建议、硬负向挖掘、集成和网络大大提高了跟踪的性能。

大多数现代的跟踪器，使用一个矩形锚框来初始化目标，并且估计他在随后的帧中的位置，此方法很难正确地表达出对象的框。

SiamMask起源于一个superpixerl-ed方法，但是SiamMask可跟中的频率比该方法高。当使用CNN特征时，superpixerl-ed方法的速度会下降六十倍，下降到0.1fps。

**半监督视频对象分割**

对象跟踪的基准：跟踪器接受连续的视频帧

多数的方法经常注重在获取一个可以超过典型视频帧率的速度，相反地，半监督**VOS**算法更加关心对象的准确代表

挖掘视频帧的连续性

许多方法通过图标记方法传播了第一帧的监督分割掩码（如：使用时空马可尔夫随机场，它的临时依赖性被光流所建模，而空间依赖性被CNN所表达）

另一个受欢迎的策略是独立处理视频帧（如 ：OSVOS-S,没有利用任何临时信息，而是事先训练好的依赖全卷积网络来分类，然后在测试的时候使用第一帧的真实掩码来微调。MaskTrack是从零开始对单个图像进行训练，但它确实利用了测试时的某种形式的时间性，使用最新的掩模预测和光流作为额外的网络输入）。

以准确率为目标VOS方法经常在测试时使用微调、数据增强、光流等技术，但是只适合低速率和非在线（DAVIS需要很长时间才能运算完成）

视频分割更快的方法（这些方法都比SiamMask差）：

使用元网络“调制器”在测试期间快速调整分割网络的参数

不使用任何微调，并且使用一个在许多场景中训练出来的编码解码的Siamese结构

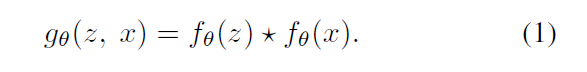
**方法**

使用全卷积的Siamese框架

**参考的全卷积Siamese框架**

**SiamFC**

将图像z（从给的图像中切割出来的目标，**w\*h crop**）与一个搜索图像x（即视频帧中的目标对象）进行比较，得到一个密集的响应图（**ROW**）



**g^nθ(z,x)编码样本z和x的第n次训练的候选窗口的相似性**

两个输入来自同一个CNN fΘ，生成两个互相关联的特征图

使用**logistic loss**来训练

**参考SiamFC的地方**：

SiamMask参考了SiamFC的此公式，SiamFC的目标是用响应图的最大值来联系目标的在区域x中的位置。然而，为了使ROW能够编码目标对象更加丰富的信息，SiamMask将简单的互关联替换为深度方向的互相关，并且生成一个多通道的响应图。

使用**Lsim loss**来训练

**SiamRPN**

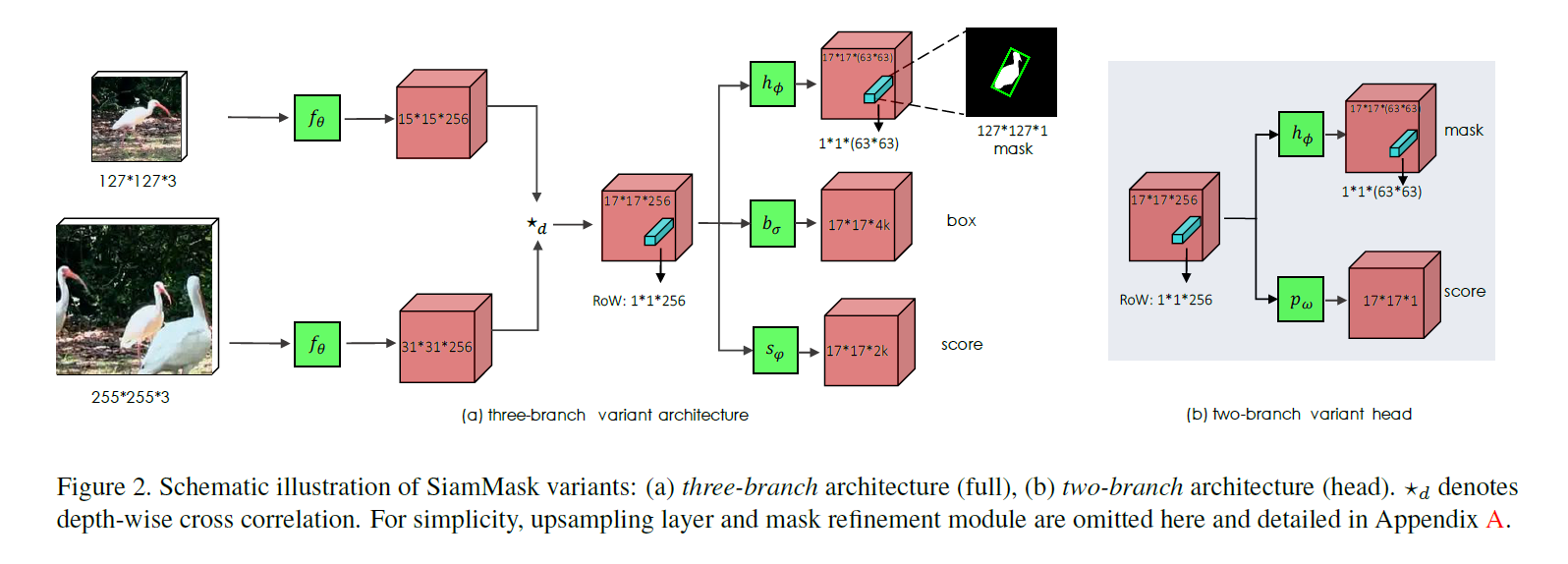
使用一个区域建议神经网络（faster RCNN）**RPN,**该网络使用可变长宽比的边界框来估计目标位置

在SiamRPN中，每个Row编码一组k个锚框和相应的目标/背景分数。因此，SiamRPN的输出框预测带有分类的分数。

两个输出的分支分布用 smooth L1 loss、交叉熵损失。

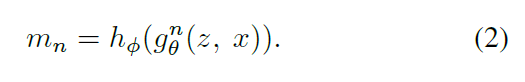
**参考SiamRPN**

**将两个损失分别参考为Lbox和Lscore**



**SiamMask**

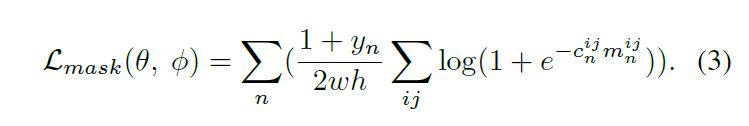
与现存的跟踪方法依赖低保真对象表示不同的是SiamMask讨论了每一帧的二进制分割掩码的重要性。除了相似分数和方框，这对全卷积的Siamese网络来说，可以编码必须的信息来得到一个像素级别的二进制掩码，从而得到额外的分支和loss



使用一个简单的两层的带有学习参数ф的**神经网络hф**来预测w\*h的二进制掩码。设mn代表第n次预测得到的二进制掩码。

**Loss function**

训练期间，每一个**ROW**被一个带有真实值二进制标签yn∈{±1}所标志，并且跟一个真实值掩码Cn(w\*h)所关联。C^(ij)n ∈ {±1}表示在第n次候选ROW的对象掩码像素（i，j）的标签



掩码预测任务的loss：Lmask是一个在所有的Row上的二进制的逻辑回归

hФ的分类层由w\*h的分类级组成，每一个都明示了一个给定的像素是否属于候选窗口的对象。Lmask只考虑有用的Row，即 yn == 1

**Mask representation**

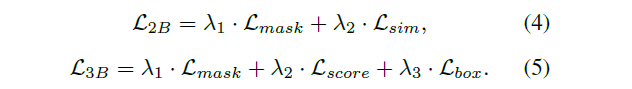
与FCN和Mask R-CNN的语义分割不同，SiamMask对象的扁平代表开始生成掩码，这个掩码代表与一个（17\*17）的RoW对应，Row由fθ(z)和fθ(x)之间的深度相关产生。

分割任务的网络hΦ由两个1\*1的卷积层组成，其中一个带有256通道而另一个有63^2个通道。这使得每一个像素分类器可以使用包含在**整个RoW中**的信息，并且因此可以获得一个完整的视野在x中和它相关的候选窗户上，这可以消除短距离内看起来像是目标的干扰的影响。

**使用上采样层和跳过连接组成的多重提炼模块来合并低和搞分辨率的特征**

**两个变体**

将SiamFC和SiamRPN的架构增加了分割分支和loss Lmask,得到二分支和三分支，这两个分支充分利用了多任务loss L2B和L3B



L3B：如果一个Row的锚框和真实锚框之间有至少0.6的IOU（当衡量两个边界框的相似度时，我们通常将Jaccard系数称为交并比（Intersection over Union，IoU），即两个边界框相交面积与相并面积之比，如图9.2所示。交并比的取值范围在0和1之间：0表示两个边界框无重合像素，1表示两个边界框相等。），那么yn=1.否则yn=-1.

L2B:部搜寻等式4和5的超参数，简单地设置lada1=32，lada2=lada3=1.

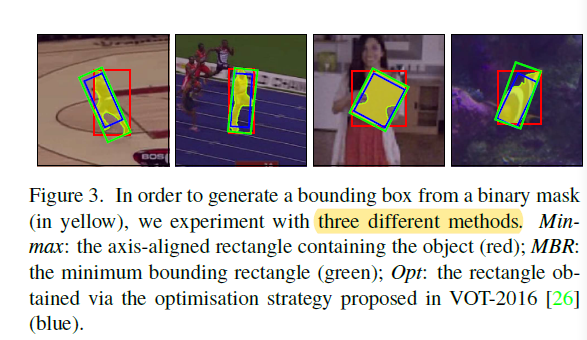
box和score输出的任务特定分支由两个1\*1个卷积层组成。

**方框的生成**:三种策略

**Min-max（red）:包含对象的轴向矩形**

**MBR（green）: 最小边界矩形**

**Opt(blue):VOT-2016提出的优化策略得到的矩形**



**补充细节**

**网络架构**

两个变体都使用ResNet-50直到第四阶段的最后卷积层作为我们的脊梁 fθ

为了载深层次获得高空间的分辨率，使用带stride为1的卷积层减少输出的stride到8

使用扩张的卷积来增加善于接受的区域

在模型中，增加一个不可分享的调整层(1\*1 conv with 256 outputs)给可分享的fΘ

**训练**

同SiamFC一样，使用 examplar和search image patches ，分别为127\*127和255\*255像素

训练期间，随机抖动examplar和earch pathches

考虑随机平移（up to ±8像素) 和 尺度条件(2^(±1/8)和2（±1/4）分别给examplar 和 search)

网络主干预训练在ImageNet-1k分类任务上，在头5个epoch使用带有一个首先预热阶段的SGD，他的学习率从10^-3线性增加到5\*10^-3 ，然后对数递减直到5\*10^-4在接下来的15个epoch。

训练集：COCO、ImageNet-VID、YouTube-VOS

**推论**

在跟踪期间，SiamMask只是评估每一帧，没有任何调整

在我们的变体中，我们使用在分类分支中获得最大分数的位置来选择输出掩码

在应用了每个像素的sigmoid之后，我们将掩模分支的输出在0.5的阈值处进行二值化

在二分支变体中，在第一帧视频后的每一帧，给mask配上Min-max box ，并且使用它作为参考来切割下一帧的搜索区域。

在三分支变体中，发现更加有效地来利用box分支的高分output作为参考

**实验**

**在vot-2016和vot-2018上评估**

都用旋转范围框来标注

使用vot-2016来理解不同形式的代表如何影响算法的表现

在第一次实验中，使用IOU和 Average Precision（AP）@{0.5，0.7}IOU

使用官方的vot工具包和预计平均重叠（EAO），与vot-2018的表现最好的比较。EAO是一个同时考虑accuracy和robutness的措施

**对象表示的重要性**

现存的跟踪方法通常用一个混合的或者可变的长宽比预测轴向边界框。关注每帧生成一个二进制掩码可以再多大程度上改进跟踪。为了专注在代表准确性上，忽略时间方面并且随机采样视频帧。接下来的方法从vot-2016的序列上，随机裁剪的搜索补丁上测试（在±16像素内转化和在2^(1±0.25)的规模内变形）。

作为参考引入SiamFC和SiamRPN作为固定和可变长宽比的代表方法，和三个获取每帧真实信息并用作不同表示策略上限的Oracle

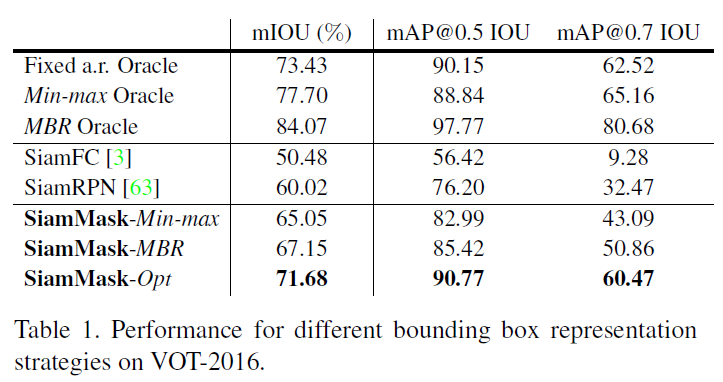
固定长宽比oracle：使用每一帧的真实区域和中心位置，但是给其中一个第一帧修改长宽比并且生产一个轴向的box

Min-max oracle使用真实bounding box的最低范围包围矩形来生成一个轴向的bounding box

MBR oracle使用旋转的最小边界矩形的真实值

SiamMask-Opt虽然有最高的IOU和mAP，但是由于它缓慢的最优化过程，它需要很大的计算。

SiamMask-MBR：显示了如何从对象的二进制掩码中获得一个旋转的边界框



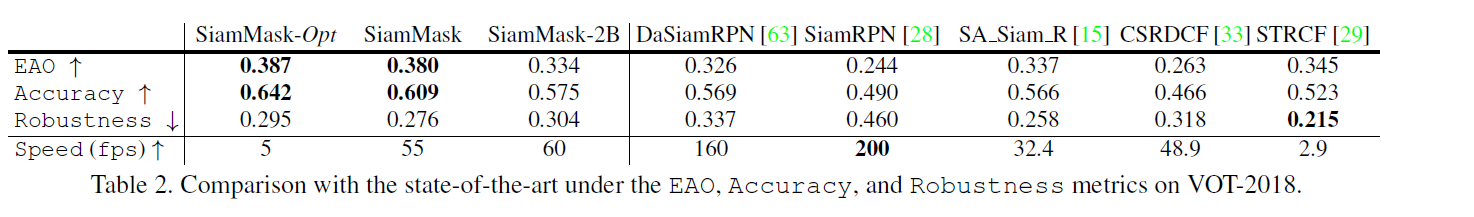
**在VOT-2018和VOT-2016上的结果**

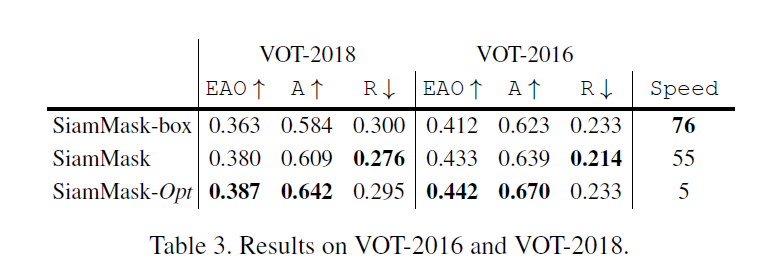
用MBR策略的二变体SiamMask和SiamMask-Opt和五个最近发表的最优跟踪器在VOT-2018上比较

三分支变体比DaSiamRPN更胜一筹，当一秒跑55帧得到0.380的EAO，即使没有box regession分支，更简单的两个分支变体（SiamMask-2B）获得0.334的高EAO

SiamMask-Opt 从二进制掩码寻找最佳的旋转矩形带来最佳的整体表现，但是付出巨大的计算代价

算法在accuracy上比基于相关过滤的跟踪器CSRDCF、STRCF表现优越，因为SiamMAsk依靠一个更加丰富的对象表现





**在半监督VOS上的评估**

模型可以被使用在VOS上来获得有竞争性的表现，担不需要再测试的时候有任何的调整

和典型的VOS方法不同，SiamMask可以在线计算，实时运行并且只需要一个简单的bounding box

数据集和设置

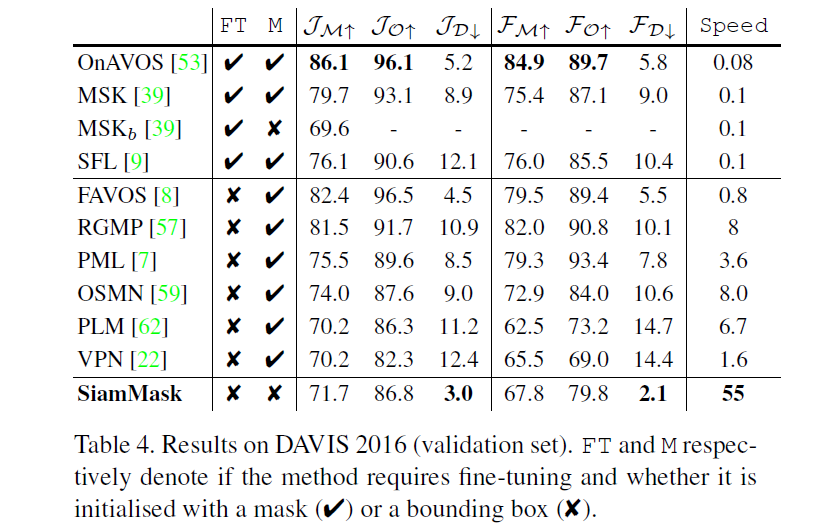
在DAVIS-2016和DAVIS-2017和YouTube-VOS标准上的表现

对每个DAVIS数据集，使用官方的表现措施：Jaccard index 来表达区域相似性和F-measure来表达轮廓accuracy。

对每个措施C∈{J，F}，三个统计学数据：mean Cm，recall Co和decay Cd，告诉我们随着时间的推移性能的增加/减少。

对YouTuVOS，报告了mean Jaccard index 和 F-measure给两个可见的（Js，Fs）和不可见的的策略（Ju，Fu），O是四个措施的平均值

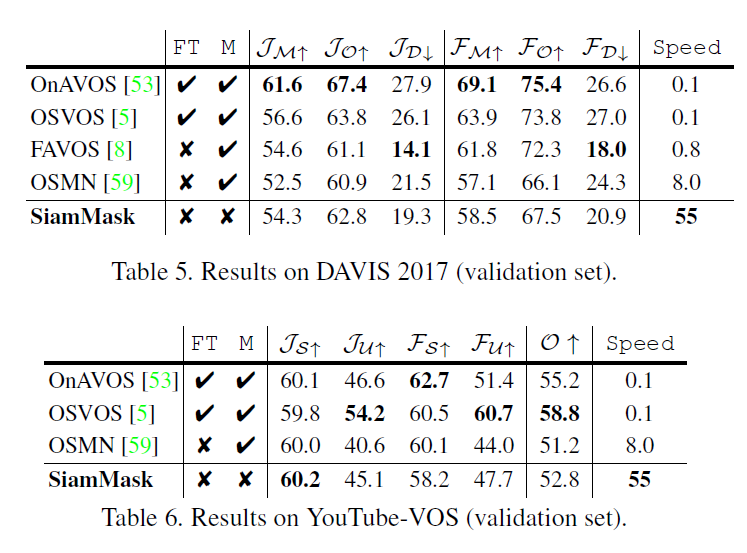
从第一帧提供的掩码中提取轴向包围框来初始化SiamMask，与大多数VOS方法类似，在同一个视频中有多个对象。



**在DAVIS和YouTube-VOS上的结果**

在半监督配置中，VOS方法用二进制掩码进行初始化，当中许多的方法都是需要在测试时密集计算的技术，如微调，数据分割，MRF/CRF上的干涉和光流。

因此，VOS通常需要几分钟来处理一个短序列。很明显，这些策略使得在线应用不可能，由于这个原因，在比较中主要聚焦在快和表现最好的方法上



**深入分析**

网络结构

AN和RN表示我们是否使用AlexNet或ResNet-50作为共享主干，当带有"w/o R"，意味着我们的方法没有使用Pinheiro的提炼策略

更新fΘ的架构就可以获得一个重要的改善，但是需要消耗速度，特别是SiamRPN

SiamMask-2B和SiamMask在他们的baselinesSiamFC和SiamRPN上有很大的改善

Pinheiro 的细化方法对轮廓精度FM非常重要，但对其他指标的精度则相对较低。

多任务训练

在推理期间修改两个变体，他们分别从score branch的轴向（SiamMask-2B-score） 或者 box branch（SiamMask-box）报告一个bounding box

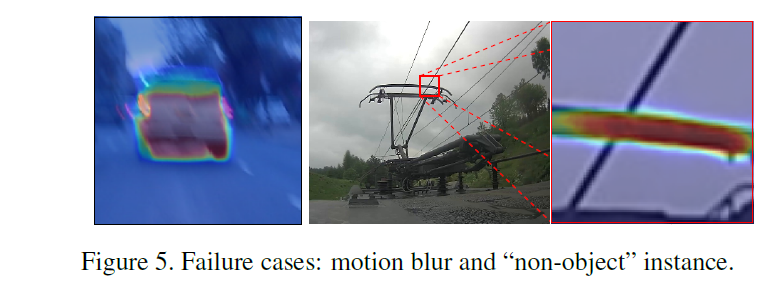
在推理期间没有使用mask branch

最高的计算量来自fθ

失败的场景

运动模糊

“非对象”实例



**结构细节**

**网络主干**

对每个变体，使用一个ResNet-50直到第四个阶段的最后一个卷积层

为了在深层得到更高的空间分辨率，通过使用stride为1的卷积减少输出的stride到8

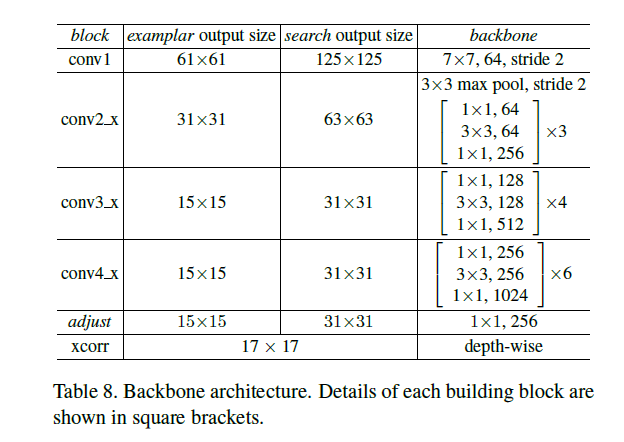
使用空洞卷积来增加可接受的区域,conv4\_1的conv3\*3层将stride设置为1 ，将dilation rate设置为2

跟原生ResNet-50不同的是，在conv4\_x没有缩减像素采样

给主干增加了一个调整层（一个1\*1的卷积层，带有256的输出通道）

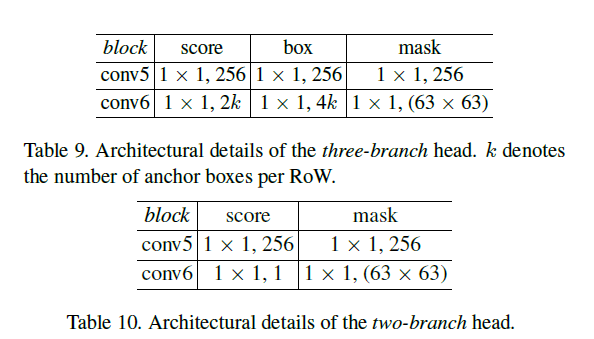
examplar和search patches 共享网络的参数，参数来自conv1到conv4\_X，但是adjust 层的参数是不共享的

调整层的输出特征按深度相关联，得到大小为17\*17的特征图



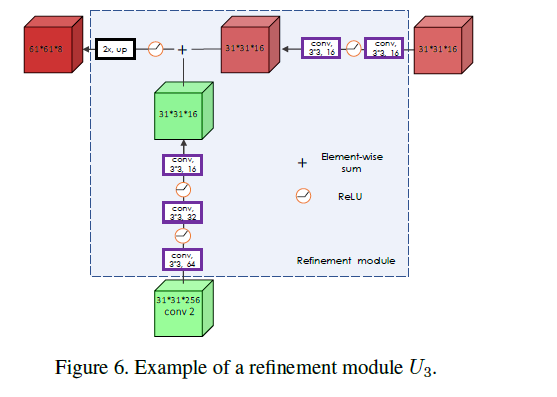
**网络头**

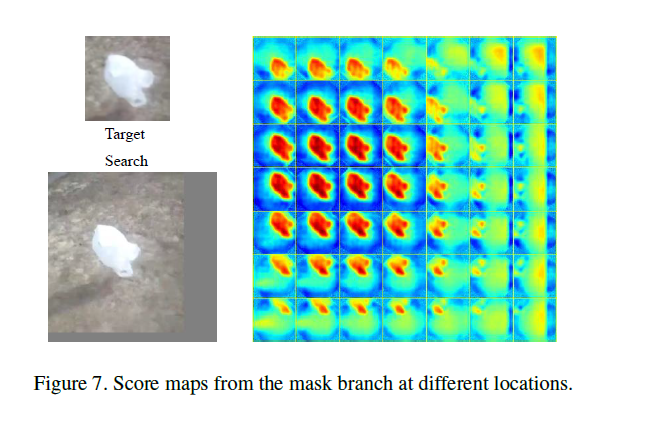
conv5 block包含一个规范化层和非线性的ReLu，但是conv6只是有一个1\*1的卷积层组成



**掩码细化模块**

为得到一个更加精确的对象mask，使用由上采样层和跳过连接组成的多个细化模块合并低分辨率和高分辨率特性





**进一步定性结果**

不同的位置的不同掩码

模型给每个RoW生成一个mask

在推理阶段，依赖socre branch来选择最后的output mask（使用包含最大score的位置）

图7的示例说明了mask分支生成的多个输出掩码，每个掩码对应于不同的Row

