### 基于卷积神经网络的目标跟踪算法综述

#### 1 引言

##### 1.1研究背景

目标跟踪是计算机视觉领域的一个重要分支，是模式识别，图像处理，计算机视觉，机器学习等学科的交叉研究，有着广泛的应用，如视频监控，虚拟现实，人机交互，图像理解，无人驾驶等。

目标跟踪的通常任务是在视频的第一帧给定一个目标的矩形框，然后后续这个矩形框紧跟着要跟踪的物体，在给定非常有限的信息的视频中定位运动物体。而目标跟踪算法的主流则是从传统的特征提取加机器学习到现在的基于神经网络的深度学习，自从2012年AlexNet问世以后，计算机视觉的各个领域都有了巨大变化。目标跟踪算法一直以来都面对着的外界因素、目标外观变形等难点与挑战，在卷积神经网络技术(Convolutional Neural Network)强大的特征表征能力加持下，都得到了不同程度的解决,因此基于CNN的目标跟踪算法逐渐成为研究的主流技术。

##### 1.2 研究难点

目标跟踪领域的难点问题主要来自于 2 个方面: 首先是目标在模型跟踪过程中一直处于变化的状态，主要包括姿态、形状和尺度的变化带来的特征变化问题； 其次，外部环境的变化也会带来一系列挑战，主要包括光照变化、运动模糊、遮挡、背景干扰等问题。这两方面的问题共同为在线目标跟踪任务带来了很多困难。其中姿态、形状和尺度的变化会导致同一目标在不同帧中的表观特征出现较大的变化，从而为目标特征匹配带来麻烦; 而外部环境的变化会导致表观模型学习到很多干扰信息，所以在比较复杂的外部环境条件下跟踪算法很难学习到目标和背景之间明确的辨别性信息，从而导致跟踪结果容易发生偏移。

##### 1.2 研究发展

跟踪算法可以被分为两类，分别是生成模型方法、相关滤波算法，基于相关滤波方法属于判别模型方法这一类别。传统的相关滤波方法的优势是速度快，可以在 CPU 上实时运行，但精度一般【11】。随着深度网络的出现与发展，计算机视觉研究者在应用CNN的基础上提出了不同的解决思路。

**Pic 2-1 目标跟踪基本流程**

目标跟踪的基本流程主要分为 3 个部分：外观建模部分, 搜索策略部分, 模型更新部分，如Pic 2-1所示。 目前的跟踪算法可以将其分为两类，一类是生成式，另一类是判别式，这和机器学习算法的分类有相似之处。

生成式算法通过描述目标的表观特征，处理当前帧与下一帧的关系。主要包括稀疏编码（sparse coding）、在线密度估计（online density estimation）等。它的缺点是过于关注目标本身，忽略背景信息，容易产生漂移现象。

判别式算法区分目标和背景，表现比生成式算法更为鲁棒，判别式模型将跟踪问题看做分类或者回归问题，目的是寻找一个判别函数，将目标从背景中分离出来，从而实现对目标的跟踪。

#### 2 经典的目标跟踪算法

##### 2.1 生成式建立模型的方法

生成类方法是在当前帧中对目标区域建模，下一帧寻找与模型最相似的区域，该区域就是预测位置，此类方法经常使用卡尔曼滤波【1】、粒子滤波【2】、均值漂移【3】等算法。

卡尔曼滤波的特点是对目标的运动模型进行建模从而估计目标在下一帧的位置，不对目标特征进行建模。杨鹏生等人【4】提出了一种基于改进扩展卡尔曼滤

波的目标跟踪算法，通过构建时间差和信号到达方向的观测方程，利用几何和代数关系化简得到伪线性模型对目标运动轨迹进行跟踪，取得了优于传统卡尔曼滤波算法的效果。粒子滤波基于粒子分布统计，寻找一组在状态空间中传播的随机样本来近似的表示概率密度函数，用样本均值代替积分运算，进而获得系统状态的最小方差估计。刘芳等人[12]针对复杂背景变化和目标尺度变化等跟踪中常见的问题，提出了一种基于深度特征和模板更新的自适应粒子滤波目标跟踪方法，解决由目标模板更新精度低导致跟踪算法鲁棒性差的问题。均值漂移方法基于概率密度分布，沿着概率梯度上升的方向，迭代收敛到概率密度分布的局部峰值上。

Tomas 等提出的生成模型方法 ASMS 算法[13]在经典的均值漂移框架下进行了尺度估计的理论推导，并引入了正则项来解决两个问题：由背景混乱引起的尺度扩张和自相似物体的尺度内爆。还有一类经典算法是基于特征点的光流跟踪算法，对目标物体提取特征点，在下一帧计算特征的光流匹配点，进行统计从而得到目标位置。王忠民等人[14]针对物体快速移动以及均值漂移算法误差累积造成的目标漂移问题，提出了一种融合目标检测算法YOLO【15】与均值漂移的目标跟踪算法，进行有效的跟踪。

生成式模型通常寻找与目标模板最相似的候选作为跟踪结果，这一过程可以视为模板匹配。而判别式模型通过训练一个分类器去区分目标与背景，选择置信度最高的候选样本作为预测结果。由于使用各种机器学习方法所以判别类方法普遍都比生成类好。

##### 2.2 相关滤波方法

2010 年后，基于相关滤波的算法开始流行，其特点是算法运算速度快。传统相关滤波器的算法利用循环相互的特性，在傅里叶域中进行运算，此类算法有效更新过滤器的权重从而实现在线跟踪。初期的相关滤波算法使用浅层特征主要由灰度特征、纹理特征、颜色特征等组成。

Bolme 等提出的 MOSSE【16】使用单帧初始化时产生稳定的相关滤波器。对于光照、缩放、姿势和非刚性变形的变化时，基于 MOSSE 滤波器的跟踪器具有

鲁棒性，同时以 669帧的速度运行。Henriques等提出的CSK【17】算法使用了一种基于循环矩阵的核跟踪方法，并且从数学上完美解决了密集采样的问题，利用傅里叶变换快速实现了检测的过程。算法通过使用高斯核计算相邻两帧之间的相关性，取响应最大的点为预测的目标中 心 。 Henriques 等 人 在 CSK 算 法 的 基 础 上 提 出 了KCF/DCF 算法【17】，这是 CSK 算法的多通道特征改进版

本。CSK 中使用的是灰度特征，KCF 使用 HOG 特征和高斯核函数，DCF 使用 HOG 特征和线性核函数。在此之后，Danelljan 等提出的 CN【20】算法是 CSK 算法的多通道颜色特征改进算法，而该团队的 DSST【21】方法只用了 HOG 特征并且专门训练类似 MOSSE 的相关滤波器检测尺度变化，开创了平移滤波和尺度滤波相结合的方法。关于颜色特征，吴晓光等人【22】在基于空时上下文和核函数循环结构的目标跟踪器的基础上，介绍基于颜色特征的跟踪器，并针对其在目标被遮挡、尺度变化和光照发生变化时易发生跟踪漂移的问题，提出自适应学习速率和自适应高斯核尺度因子两种方法，分别对训练模型的更新和标记进行改进，减少目标模型累积错误，提高跟踪过程准确性。

|  |  |  |  |
| --- | --- | --- | --- |
| Tracker | accuracy\* | | FPS |
| KCF | 0.15 | 172 | |
| DCF | 0.14 | 292 | |
| CSK | 0.11 | 40 | |
| TLD | 0.14 | 28 | |
| CT | 0.13 | 64 | |

\*注: 成功率为LaSOT测试数据下Overlap threshold等于0.6 的评估结果

**Table2-1 几种经典跟踪算法测试结果**

#### 3 基于CNN的目标跟踪算法介绍

 卷积神经网络( CNN ) 是一种前馈神经网络，其中每个神经元的响应与前一层感受野范围内的神经 单元相关 。 卷积神经网络通常包含卷积层、激活函 数、池化层以及全连接层。其中，卷积层由若干卷积单元组成，卷积运算是为了对输入进行特征提取，浅层网络可能是提取一些低级的特征如边缘、轮廓和角点等，深层网络从低级特征中提取更为复杂的特征。池化层即降采样层，通常卷积层之后会得到维度较大的特征，池化后可得到维度较小的特征。全连接层把所有局部特征结合变成全局特征。代表性的 CNN 模型有 Alex Net和Res-net。

基于CNN的目标跟踪算法是一种判别式算法，当前取得SOTA(State of the art)效果的基于CNN的跟踪算法，按照网络架构的不同大体可以分为三类，第一类是使用孪生网络架构的siamese系列算法，第二类是在"两步法"目标检测网络架构的基础上完成跟踪任务的系列算法，第三类则是采用序列化的网络网络结构，分步完成分类和回归任务。

下面介绍不同网络架构思路下的代表性工作。



**Pic 2-2 各时间节点代表工作**

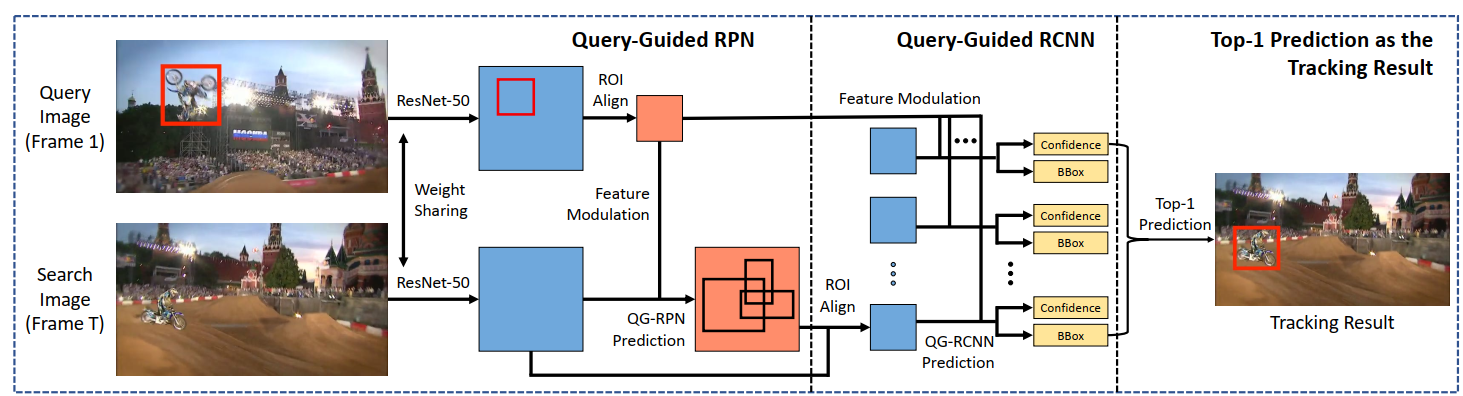
##### 2.1基于检测网络的目标跟踪算法

检测式跟踪的方法，即tracking by detection，把连续的跟踪问题看做是对单帧图片的目标检测，只是要检测的目标只有第一帧选中的“特定目标”，使用目标检测算法将每帧中与第一帧最相似的目标检测出来，得到对应的位置坐标, 可信度，同时利用目标图的信息去指导算法跟踪特定的目标。

这类算法的优势是会不对跟踪目标做时间一致性假设，在目标脱离视野或者镜头出现剧烈抖动的情况下重新捕获目标的能力强；同时检测网络对象尺寸和纵横比变化保持鲁棒性，而使用流行的互相关操作很难做到这一点在数据集上表现的精度更高。缺点是由于采用目标跟踪的多通道复杂网络结构造成信息冗余，算法的实时性较差，FPS在十数帧以下。具有代表性的工作是GlobalTrack算法和Siam R-CNN算法。

1. GlobalTrack

GlobalTrack将搜索帧特征图与目标物体特征图进行卷积操作后进行二阶段目标检测，收到了较好的效果。检测式跟踪不甚重视帧间联系，每一帧对目标的搜索相对独立，不会引入累积误差，主要克服了长时间跟踪问题，完全脱离了对模型的在线更新。但同时抗干扰物的能力较差，容易出现跟踪效果的漂移,同时速度较为缓慢。

GlobalTrack的动机是实现基于全局搜索思想的视频长期目标跟踪，它使用了一种很直接的体系结构（pic 3-3），在这项工作中，跟踪器对视频序列时间一致性（即目标状态平稳变化）的情况没有任何假设或约束，算法不进行在线学习，无位置或尺度变化的惩罚，无尺度平滑和轨迹改善，此不存在累积误差。因此，它在的性能不受先前帧跟踪效果不好的影响，这使其适合长序列的跟踪。 该方法是基于两级对象检测器开发的，它由两个组件组成：用于生成特定查询实例候选的查询引导区域提议网络（Query-Guided RPN），以及用于对这些候选进行分类并生成的查询引导区域卷积神经网络(Query-Guided R-CNN)，两者都是在**MMdetection**项目下的二次开发，是基于其RCNN、RPN组件进行的构建。它能够以单个查询为指导对任意实例执行全图像和多尺度搜索。

**Pic 2-3 GlobalTrack网络结构**

其最大的创新就是在于查询引导的RPN和查询引导的RCNN模块。跟踪的流程与两阶段的目标检测（e.g. Faster R-CNN）接近，但在输入网络前加入了Feature Modulation操作（Query-Guided），即将第一帧指定目标目标区域输入网络（e.g. ResNet-50）进行特征提取、ROI Align校齐，然后将得到的固定大小的特征图与新的搜索图像的特征图进行卷积操作。查询引导的RCNN对候选目标进行更精细的分类和定位，最后的输出结果直接使用置信度最高的候选对象的包围框作为结果。在消融实验中，作者对比了QG-RPN与RPN效果差异，加入QG后召回率、准确度都有4%以上的提升。

另外一个创新的部分是交叉查询损失，同一张图片用不同的目标框进行二阶段重追踪，得到loss后求平均值，用于提高针对干扰因素的方法的鲁棒性。

从文章对比结果来看，GlobalTrack在不使用任何后期处理等步骤的情况下，获得了SOTA结果，根据跟踪器特性，视频在评价数据集中的长度越长，GlobalTrack的性能应该越好。

2) Siam R-CNN

Siam-RCNN同样是一种两段式检测器加持的跟踪算法，论文中主要方法是在孪生两阶段重检测的基础上，引入论文新提出的基于tracklet的动态规划方法，可以充分利用对第一帧模板和先前预测帧的优势，解决追踪问题。

  论文的主要贡献是，在 R-CNN（Rich Feature CNN）网络结构的基础上，提出了一种新的基于tracklet的动态规划方法。通过动态规划的方式关联不同段的tracklet，在较大提升了跟踪的准确度的同时克服了跟踪的漂移。

文中引入了tracklet的概念，作者将一个追踪A=(a1,...,aN)视为由N个不重叠的tracklets（跟踪片段）组成的序列，tracklet ai的结束时间，肯定在下一个tracklet的开始时间之前。这也表明tracklet是一组时间上连续的检测。在长期追踪中它可以理解为，在目标的一个存续期间内的每帧检测结果（目标追踪框）的集合。而在长时追踪中，目标可以有多个存续期间，将这多个存续期间的tracklets组合起来，就是在整个视频中对这个目标的追踪A了。

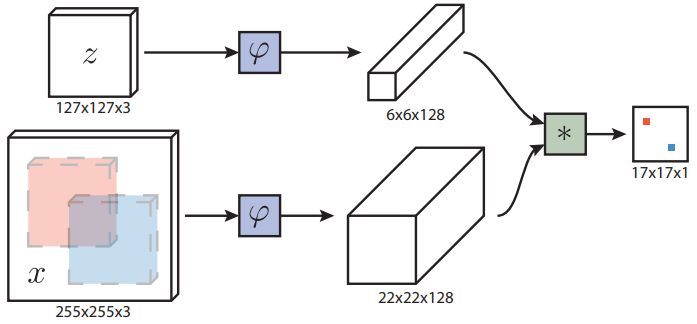
Tracklet的引入，使得该算法利用重新检测第一帧模板和先前帧预测的优势，对要跟踪的对象和潜在的干扰对象的完整历史进行建模。 从而可以做出更好的跟踪决策，并可以在长时间遮挡后重新检测被跟踪的对象。

##### 3.1 基于孪生网络的目标跟踪算法

孪生网络结构是一种特殊的神经网络架构，由两个或更多子网络构成，其特点是同时接收两个图片作为输入并且两个神经网络权值共享。孪生网络结构的主要思想是找到一个可以将输入的图片映射到目标空间的函数，使得目标空间中的简单距离近似于输入空间的“语义”距离。更准确的说该结构试图找到一组参数，使得相似度度量在属于同一类别是小的，在属于不同类别时是大的。该网络过去主要用于度量学习，用来计算图像、声音、文本等信息的相似性。尤其是在人脸验证领域上的应用【19-20】。孪生网络结构还有一个特点就是可以自然地增加训练数据量，因为每次输入一对图像。这样就可以充分利用有限的数据集来训练网络，这一点在目标跟踪领域非常重要，因为和目标检测相比，跟踪领域的训练数据集较少。

###### 3.1.1 孪生网络的应用

随着 SINT[35]和 Siam FC[36]的出现，孪生网络结构跟踪器成为了解决单目标跟踪问题的一个重要方法。基于孪生网络的算法特点是通过孪生网络学习一个匹配函数，在得到第一帧的边界框后，后续的所有帧都在采样后和第一帧中的目标进行相似度计算。由于网络模型不更新，目标即使被遮挡也不会影响后续的找回，这样可以极大提高跟踪算法的鲁棒性。Tao 等提出的 SINT 方法第一个提出了基于孪生网络结构的目标跟踪算法，该算法在图像中生成多个候选目标，经过网络后与示例图像作比较，然后选择距离最小的候选目标。即开创性地将目标跟踪问题转化为匹配问题，并通过神经网络来实现，但每次处理大量的候选框是非常耗时的操作。同年 Bertinetto 等提出的 Siam FC 方法更具有实用性。该方法训练一个深度网络以在初始离线阶段解决更一般的相似性学习问题，然后在跟踪期间简单进行在评估。该算法训练一个全卷积孪生网络，以在更大的搜索图像中定位候选图像。新的孪生结构对搜索图像进行了完全卷积：使用双线性层实现密集和高效的滑动窗口评估。该双线性层实际上计算其两个输入的互相关，将示例图像与候选图像进行比较，并且如果两个图像描绘相同的对象则返回高分，否则返回低分。为了在搜索图像中找到对象的位置，可以穷尽地测试所有可能的位置，并选择与对象的外观具有最大相似性的候选图像。



**图 2-3 SiamFC网络结构**

网络结构如图 2-3所示，由模板分支和检测分支组成。模板分支负责接收目标作为输入（用 z 表示），目标通常是第一帧给定的边界框。检测分支接收在当前帧作为输入（用 x 表示）。卷积神经网络中的两个分支共享参数。Siam FC 使用深度卷积网络作为函数，使用孪生架构的深层网络进行相似性学习。孪生网络对两个输入应用相同的变换，然后根据计算相关性，其中函数g(x,y)可以被认为是卷积操作，其中卷积核是 .

在保证精度的前提下实现了70FPS以上的跟踪速度，是基于CNN的目标跟踪跟踪算法在实时性上的突破。但Siamese FC也采用了跟踪问题中常采用的多尺度测试的方法，对精度的提升形成了限制。

###### 3.1.2 孪生网络的改进

初期的孪生网络在速度上远超实时，但精度上不如结合深度特征的相关方法，随着研究者对孪生网络的不断改进，来自其他领域的功能模块被应用在孪生网络结构中，改进的基于孪生网络结构的跟踪器使用更好的方法取代了传统的多尺度检测，并且使用了更深的网络结构，在精度和速度上追赶结合深度特征的相关滤波跟踪器。这一小节将改进从模块的引入和backbone网络的优化两方面介绍。

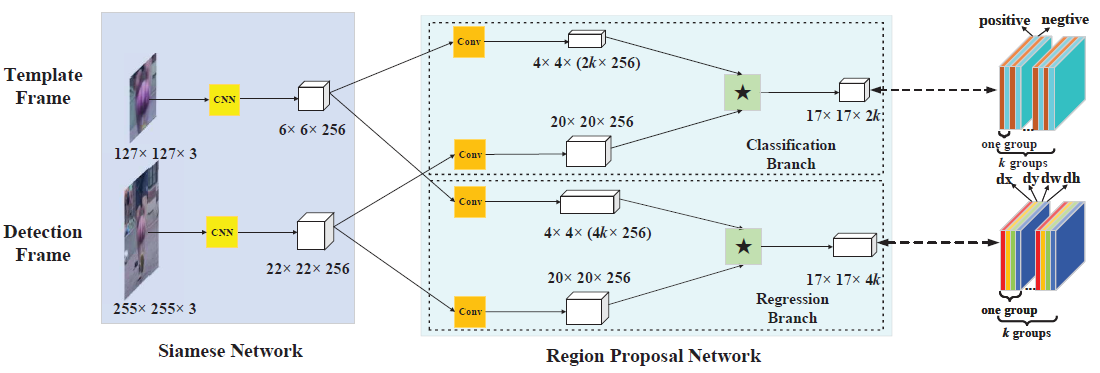
1）模块的引入

Siam FC中一个非常大的弱点是没有尺度估计。Siam FC 输入三个尺度的图，经过网络对比后选出响应最大的作为目标的位置。手工设计的尺度长宽比例固定，这一点非常不合理。为了改进目标框的精度问题，后续工作Siam-RPN，Da SiamRPN引入了RPN（Region Proposal Netdwork）网络。

SiamRPN，结合了孪生网络和区域提议网络（Region Proposal Network），采用目标检测中回归、分类分支进行输出的方式确定目标框，改善了SiamFC等算法用多尺度测试确定目标框的弊端，实现很高准确度的同时兼顾了速度。

图2-4展示了Siam-RPN的主要结构，分为特征提取孪生网络结构和RPN结构。，在 RPN 的分类分支中，模板图像和检测图像的特征图首先通过一个卷积层，该卷积层主要是对模板图像的特征图进行通道上的升维，令其维度变为检测图像的特征图维度的 2k 倍，其中 k 为 RPN 中设定的锚框数。此后，将模板图像的特征图在通道上按序等分为 2k 份，作为 2k个卷积核，在检测图像的特征图完成卷积操作，得到一个维度为2k 的分数图。该分数图同样在通道上按序等分为 k 份，得到 k 个维度为2的分数图，两个维度分别对应锚框中前景（目标）和后景（背景）的分类分数，是关于目标的置信度。

在 RPN 的回归分支中，模板图像和检测图像的特征图，都将首先通过一个卷积层，该卷积层主要是对模板图像的特征图进行通道上的升维，令其维度变为检测图像的特征图维度的 4k 倍，其中 k 仍为 RPN 中设定的锚框数。此后，将模板图像的特征图在通道上按序等分为 4k 份，作为 4k 个卷积核，在检测图像的特征图完成卷积操作，得到一个维度为 4k 的分数图。分数图同样在通道上按序等分为 k 份，得到 k 个维度为4的分数图，四个维度分别对应修改锚框坐标及尺寸的参数。



**图3-4 SiamRPN网络结构**

随后为了克服样本不均衡和抗干扰相似物的问题，siamRPN的工作组又提出了Da-siamRPN,“Da”即Distractor-aware，指的算法在SiamRPN的基础上对有语义信息的目标干扰会有察觉和处理。 相比于SiamRPN，改进的地方有两点：

1. 补充训练集

丰富样本的个数和种类，让跟踪更加鲁棒；作者发现在跟踪过程中跟踪器对有语义信息的实例分类能力较差，而对前背景分类能力很强，他认为造成这个问题的原因是跟踪过程中样本不均衡得分较高的这些false positive，  
正样本实例种类不够多模型泛化能力差SiamRPN所用的训练集ILSVRC2015-VID和Youtube-BB所包含的种类不太充分，为了提高跟踪器的泛化能力，作者在此基础上又引入了ILSVRC2015和COCO数据集静态图片，通过一系列增强手段（平移、调整大小、灰度化等）来扩充了正样本对的种类。为了提高判别能力，在负样本中，作者发现有语义的背景和类内干扰物都很少，所以增加了不同类别的困难负样本来避免跟踪结果漂移，以及增加了相同类别的难例负样本来更加关注目标的细节表达。其中难例负样本(hard negative)是容易被网络预测为正样本的proposal，比如roi里有二分之一个目标时，虽然它仍是负样本，却容易被判断为正样本，这块roi即为困难负样本，困难负样本可以提示网络去针对这些“易错点”。将困难负样本引入训练的方式是先用初始样本集(即第一帧随机选择的正负样本)去训练网络，再用训练好的网络去预测负样本集中剩余的负样本，选择其中得分最高，即最容易被判断为正样本的负样本为困难样本，加入负样本集中，重新训练网络，个人体会这种训练的结果有点类似于生成对抗网络，当对负样本判别越来越准确，得到的误判假性正样本就越来越与正样本相似。

1. 提出了一种增量学习方法

选择最优bbox的标准变化了，用f(z,x)记为求相似度的相关运算。最优跟踪目标选择为：

从上式中我们可以看出，其实就是当前位置p的特征图和模板z计算相似度之后，减去当前位置与干扰物相似度的加权和。与模板帧的匹配分数减去与干扰物的匹配分数作为最终分数。这样就进一步保证了选择跟踪的准确性，不易被类似物干扰。互相关操作是一个线性操作且计算量较大，于是作者进行了化简，使用分配率减小相关操作运算的次数，将Pk的特征图运算项提了出来。

另一个引入模块的算法是SiamFC++，主要工作是定义了一套基于anchor-free的目标估计的回归模块，同时提升了孪生网络系列算法的速度和准确度。同时阐明了之前的孪生网络追踪器存在的不合理性问题，具体在于4个方面：

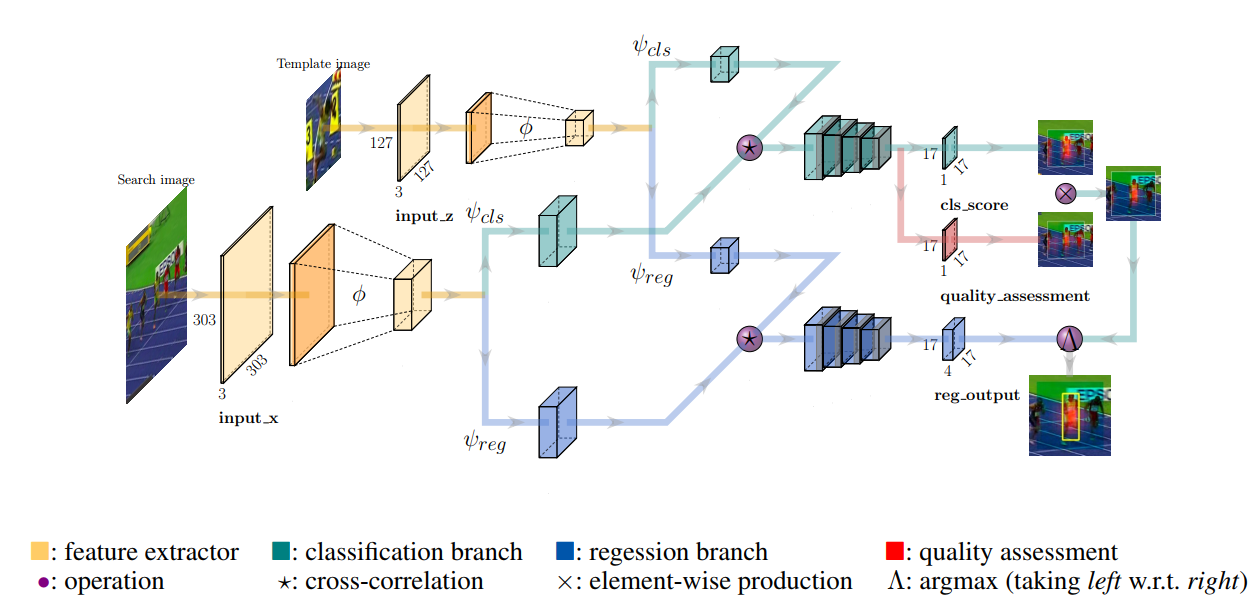
G1: 分开进行分类和状态估计。跟踪器应执行两个子任务：分类和状态估计。 没有强大的分类器，追踪器就无法将目标与背景或干扰因素区分开，这严重阻碍了目标的鲁棒性。

G2：无歧义评分。分类分数应该直接表示目标存在的置信度分数，即在“视野”中，即对应像素的子窗口，而不是诸如anchor box之类的预定义设置。 对于一个负例，对象与锚点之间的匹配（例如，基于锚点的RPN分支）容易产生假阳性结果，从而导致跟踪失败。

G3：无先验信息。跟踪方法应该没有像尺度/长宽比分布这样的先验知识，这是通用对象跟踪的目标提出的。 现有方法广泛依赖于数据分布的先验知识，这阻碍了泛化能力。

G4：估计质量评估。如之前的研究（Jiang等，2018; Tian等，2019）所示，将分类置信度用于边界框选择将直接导致性能下降。 与先前关于对象检测和跟踪的许多研究一样，应该使用与分类无关的估计质量评分。

之前应用anchor的目标是为了解决SiamFC在追踪的过程中，只能进行目标的中心位置的定位，而对于边界框的尺寸和长宽比却没有办法，这一问题。但是如果能直接回归中心到真实边界框的距离的话，就可以直接解决这一问题了。



**图 3-5 SiamFC++网络结构**

本文着重强调了质量评估分支，作者指出分类置信度与定位精度没有很好的相关性，因此有必要用额外的分支来准确的找到目标中心。实际上代表了在滑动窗中心的像素点具有高评分，远离中心点像素的评分会降低，做了一个像素的级的加权。质量评估分支的期望输出结果就是每个像素的正确PSS分数。

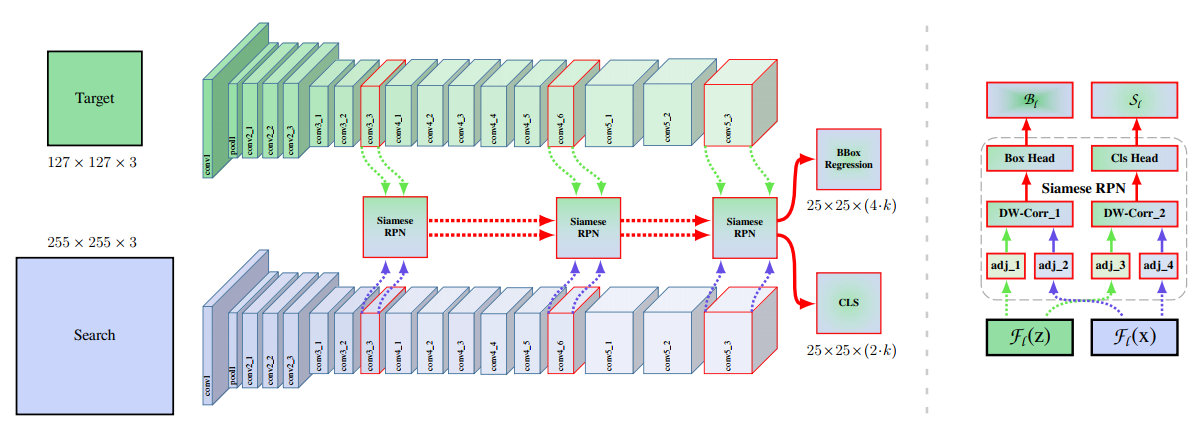
2）backbone网络的优化

Zhang 等提出的 Siam DW[44]认为随着深度学习的发展，网络层数越来越深，从而提取到更好的特征。但跟踪算法仍使用经典的 Alex Net，这是非常不合理的。论文中的实验对比了多个骨干网络，但随着网络层数的增加，跟踪效果剧烈的变差。作者分析了神经网络的多个因素，例如步长、填充和感受野。大部分的计算机视觉任务的步长经常设置为 16 或者 32。但对于跟踪人物而言，由于帧间的位移较小，步长通常取值较小，例如 4 或者 8。通常神经网络都是使用填充来确保随着层数的增加，特征图不会变得太小，并且正常采集边缘的信息。但从 Siam FC 开始，零填充被引入到网络中。在没有填充的时候，目标移动一定的

距离，特征图上原来的相应点会对应移动一定距离，这个距离和步长有关。此时特征图上响应值不变。有填充时，如果目标移动后的位置对应的感受野超出了图片的边界，就会导致响应值的改变，填充的引入导致了学习中的位置偏见。孪生网络中的感受野的设置非常重要。感受野过大，重叠就会很大，模板图像中的相邻的两个候选图像的特征会很接近，将减小定位准确度。感受野过小会降低单个特征的判别力，而神经元感受野的大量增加导致特征识别率和定位准确率的下降。作者强调网络的步长、感受野大小要在整体上同时分析。可想而知，这些因素之间并不独立，一个改变，其他的也会随之改变。作者提出了新的残差模块来消除填充的负面影响，这使得更深的神经网络结构可以应用到目标跟踪中。

基于孪生网络结构的跟踪算法将跟踪描述为目标模板和搜索区域之间的卷积征互相关操作。然而，与最先进的算法相比，孪生网络跟踪器的精度仍然存在差距，而且它们无法利用深度网络（如 Res Net-50 或更深）的特性。Li等[45]

在Siam RPN的基础上提出了Siam RPN++解答了为什么深层网络 Res Net 不能应用在孪生网络架构中，其结构如图3-6所示。



**图3-6 SiamRPN++网络结构**

孪生网络无法使用更深的网络结构的核心原因是缺乏平移不变性。作者认为原始的采样策略存在问题，该策略使得图像的中心一直有较大的权重，因此作者通过一种简单有效的空间感知采样策略，在中心进行移位，即偏移中心 16~64 个像素范围内进行均匀采样，成功地训练了一个性能显著提高的采用 Res Net 作为骨干网络的孪生网络跟踪器。此外，该方法提出了一种新的模型架构来执行逐层聚合和深度聚合，这不仅进一步提高了精度，而且减小了模型的尺寸。视觉跟踪需要从低到高、从小到大、从细到粗分辨率的丰富表示。即使在卷积网络中有深度的特征，单独的层也是不够的：将这些表示进行组合和聚合可以提高

识别和定位的能力。在 Res Net 这类深度网络中，全面卷积层获取的特征主要集中在低层信息，如颜色、形状等，对于定位是必不可少的，而缺乏语义信息；后面层的特性具有丰富的语义信息，在一些挑战场景中，如运动模糊、巨大变形时，这些信息可能是有益的。这种丰富的层次信息可以帮助跟踪。逐层聚合操作将 Conv3、Conv4 和 Conv5 输出的特征独立地输入到区域推荐网络中，由于三个区域推荐模块的输出尺寸具有相同的空间分辨率，因此可以直接对区域推荐网络输出进行加权求和。互相关计算模块是一个用来整合两个分支信息的

核心操作。作者提出了一个轻量级的互相关层，称为深度互相关，来实现更有效的信息关联，它具有更少的参数。

##### 2.3 基于CNN的序列化分类和位置估计的目标跟踪算法

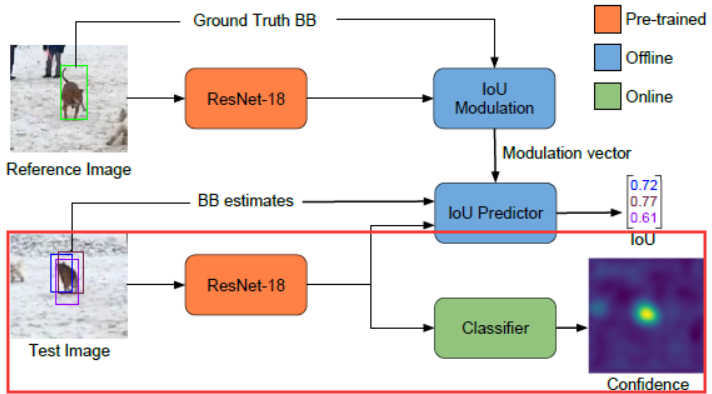
受DCF（Discriminative Correlation Filter）和IoU-Net的启发，以ATOM（Accurate Tracking by Overlap Maximization ）为代表的算法，通过序列化的分类和位置估计来跟踪目标[1]。对通过分类获得的目标的粗略初始位置进行迭代以提升精度，进行准确的框估计，该方法在准确性上产生了显着的改进，但是也带来了沉重的计算负担，每帧会随机初始化多个初始目标框，迭代细化会大大减慢ATOM的速度。此外，ATOM引入了许多其他需要仔细调整的超参数[2]。

1）ATOM

ATOM具体的目标粗定位(分类)和形状估计两个子任务模块如Pic 3-2。红框外的是一个离线训练的目标形状估计模块，红框内是一个在线训练的目标粗略定位模块，这两个任务被融合进了一个统一的网络结构，后面分开详细介绍。

作者虽然没有采用siamese网络，仍然采用了对称的特征提取网络，以利用好给定目标的参考帧的信息，上下分支特征提取的backbone网络（Resnet-18）参数是预训练且固定且共享的。

作者直接应用了ECCV18的IoUNet，bb回归的方法，给一个粗略的目标状态，，返回一个更加准确的目标状态（红框外的部分）。

reference branch通过resnet-18作为backbone提取特征，然后对目标区域

**Pic 3-2 ATOM网络结构**

进行PrPool (Precise RoI pooling)，最后通过全连接层编码成两个调制向量modulation vector，实际上是赋予权重的。这里只需要进行一次，也就是初始化的时候需要，后面推断的时候就只需要用到保存下来的modulation vector了。

test brach也是先用resnet-18对当前帧提取特征，而后对bounding box进行PrPool得到固定尺寸的特征图。这里的bounding box是classifier那边得来的一个大概的框。

使用reference branch生成的两个vector分别对两个特征图进行channel-wise multiplication，也就相当于channel attention。

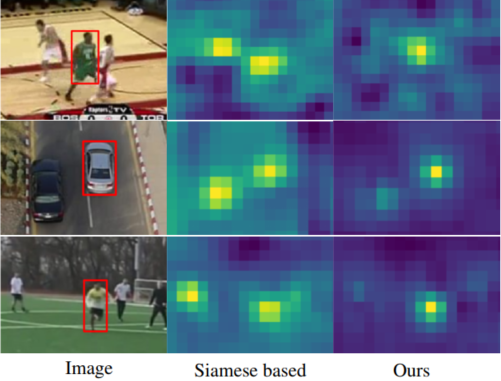
通过几个全连接层得到预测的IOU置信度。对IOU求bbox的梯度，经过几次迭代，调整bounding box最大化IOU，得到精细的预测框。

这里的classifier是一个两层的轻量级全卷积网络，通过这个分类头，输出一个置信度图，即可确定目标的大致位置。这里的classifier是在线训练的，也就是在做tracking的时候才临时训练的。因为tracking是一个target-specific的任务而不是class-specific的任务，他没有预定义的类别，而只有实体的概念，跟踪的类别是可以任意的。就算你使用了大量的类别来离线训练这个头部，也是不可用的，因为图像中可能包含多个有语义信息的实体，这样的分类器是没有办法区分开这些实体的。所以我们需要使用在线训练的方法，在tracking的时候来临时训练一个分类头部，用于输出目标物体的响应。

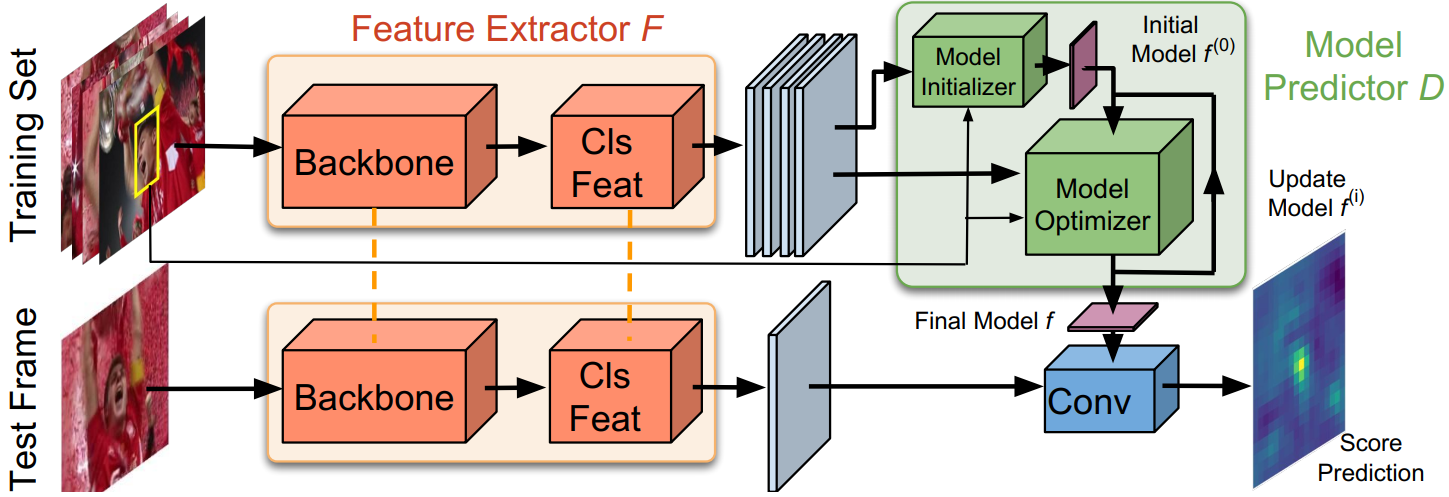
2)DiMP

另一个工作采用类似序列化分类和估计的工作是DIMP（Discriminative Model Prediction for Tracking）。

该作的论文中指出Siamese的网络框架没有考虑背景的影响，导致网络判别力不足。因此DiMP在引入新的hinge-loss损失函数和在线跟踪的机制，提升跟踪器的判别性。改进的焦点是分类网络，使分类网络产生的得分图更具有判别力。在目标周围的相似物得分被压制，排除模棱两可的相似目标。给定带有注释的第一帧，作者采用数据增强策略来构建包含15个样本的初始集合。然后使用判别模型预测架构获得目标模型。对于第一帧，用10次梯度下降来学习。只要有足够的置信度预测目标，方法就可以通过向添加新的训练样本轻松地更新目标模型。通过丢弃最早的样本，保持最新的50个样本。在跟踪过程中，通过每20帧执行两次优化器递归完成更新，或在检测到干扰波峰时执行一次递归来更新目标模型。



**Pic 3-3 DiMP可视化改进效果**

同时，作者直接还直接应用了IoUNet中boundingbox回归的方法，给一个粗略的目标状态，，返回一个更加准确的目标状态。等于说dimp只在目标分类（初定位）做了改进。

**Pic 3-4 DiMP网络结构**

reference branch通过resnet-18作为backbone提取特征，然后对目标区域进行PrPool，最后通过全连接层编码成两个调制向量modulation vector，实际上是赋予权重的。这里只需要进行一次，也就是初始化的时候需要，后面推断的时候就只需要用到保存下来的modulation vector了。

test brach也是先用resnet-18对当前帧提取特征，而后对bounding box进行PrPool得到固定尺寸的特征图。这里的bounding box是classifier那边得来的一个大概的框。通过几个全连接层得到预测的IOU置信度。对IOU求bbox的梯度，经过几次迭代，调整bounding box最大化IOU，得到精细的预测框。

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Tracker | Success Rate\* | | | FPS | | category |
| GlobalTrack | 0.52 | 6 | | | | tracking by detection |
| Siam R-CNN | 0.67 | 4.7 | | | tracking by detection | |
| SiamFC | 0.35 | 86 | | | | Siam- |
| SiamRPN | 0.38 | 102 | | | | Siam- |
| DaSiamRPN | 0.42 | 57 | | | | Siam- |
| SiamFC++ | 0.54 | 90 | | | | Siam- |
| ATOM | 0.51 | | 34 | | | Serialized cls&reg | |
| DiMP | 0.60 | | 40 | | | Serialized cls&reg | |

\*注: 成功率为LaSOT测试数据下Overlap threshold等于0.6 的评估结果

**Table3-1 基于CNN的跟踪算法测试结果**

#### 3、当前主要研究方向

3.1在线更新机制

在线更新机制（Online Update Mechanism）：为了捕捉目标( 和背景) 在跟踪过程中的变化，目标跟踪需要包含一个在线更新机制，在跟踪过程中不断更新外观模型，ATOM算法即采用了这一策略。常见的外观模型更新方式有模板更新、增量子空间学习算法及在线分类器等。但如何设计一个合理的在线更新机制，既能捕捉目标( 和背景) 的变化又不会导致模型退化、引入误差，仍是目标跟踪研究的一个关键问题。

3.2长时间跟踪

“长时间跟踪”是跟踪算法要解决的一个主要问题，大于2000帧的长期目标跟踪意味着需要算法实时计算，在追踪中途物体可能会消失再出现，而且随着光照、背景的变化和由于偶尔的部分遮挡，物体在像素上体现出来的“外观”可能会发生很大的变化。单独使用追踪器较难胜任这样的任务，因此往往结合检测器来完成任务，当然也出现了只使用检测器完成任务的GlobalTrack算法。把追踪器和检测器更好的结合使用，提升长期跟踪的准确度，是解决这类问题的主要思路。

##### 3.3 数据问题

虽然基于CNN的深度学习类算法模型在图像表观建模上有着很大的优势，但是通常深度学习模型的获取需要大量的训练数据，而这些训练数据的获取是要花费大量代价的，所以针对这个问题的研究还值得深入，目前有研究人员从少样本学习的方法中去尝试，这也是深度学习领域普遍面对的瓶颈之一。除此之外，在基于深度学习方法的目标跟踪模型中引入注意力机制也是一个热门研究方向，因为目标跟踪任务中常常需要突出目标搜索区域的中心位置，并且弱化背景对模型响应计算的干扰，而注意力机制就可以做到这一点。

### 附录

#### 单目标跟踪常用数据集及其评价指标

单目标跟踪常用的数据集主要用于测试评估算法和训练网络，其中一些包含视频序列属性全面但体量不大，仅供测试使用，如OTB和VOT数据集；另外一些数据集视频序列丰富，用于测试和训练均可，如LaSOT，GOT-10K。此外，目标跟踪训练阶段往往是两或三帧图像输入网络进行训练，为了提供丰富的负样本对和对象类别，往往会使用目标检测数据集如COCO。

**1、OTB数据集（Object Tracking Benchmark）**

1）简介

OTB 是一个评估数据集，它包含一些 benchmark 结果（即一些典型算法的测试结果），打好标记（即ground-truth.txt )的数据集，以及用来测试算法的代码库。

数据集包括50帧的序列和100帧的序列，其中50帧序列的数据集是2013年提出来的，100帧的数据集是2015年提出来的，所以OTB50也叫OTB2013，OTB100也叫OTB2015。里面涉及到灰度图像和彩色图像，也涉及到目标跟踪的11个属性，包括光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、出视野、背景干扰、低像素。

每个图像序列都对应着两个或多个属性，每个序列都对应着一个txt文件，记录着人工标注的目标中心位置和目标的大小。

2）评价指标

**精确图 precision plot**

追踪算法估计的目标位置（bounding box）的中心点与人工标注（ground-truth）的目标的中心点，这两者的距离小于给定阈值的视频帧的百分比。不同的阈值，得到的百分比不一样，因此可以获得一条曲线。

该评估方法的缺点：无法反映目标物体大小与尺度的变化。

**成功率图 Success Plot**

首先定义重合率得分（overlap score，OS），追踪算法得到的bounding box（记为a），与ground-truth给的box（记为b），重合率定义为：OS = |a∩b|/|a∪b|，|·|表示区域的像素数目。当某一帧的OS大于设定的阈值时，则该帧被视为成功的（Success），总的成功的帧占所有帧的百分比即为成功率（Success rate）。OS的取值范围为0~1，因此可以绘制出一条曲线。

**OPE**

用ground-truth中目标的位置初始化第一帧，然后运行跟踪算法得到平均精度和成功率。这种方法被称为one-pass evaluation (OPE)。

**鲁棒性评估（SRE，TRE）**

通过从时间（temporally，从不同帧起始）和空间（spatially，不同的bounding box）上打乱，然后进行评估。可以分为：temporal robustness evaluation (TRE) 和 spatial robustness evaluation (SRE)。

在一个图片/视频序列中，每个跟踪算法从不同的帧作为起始进行追踪（比如分别从第一帧开始进行跟踪，从第十帧开始进行跟踪，从第二十帧开始进行跟踪等），初始化采用的bounding box即为对应帧标注的ground-truth。最后对这些结果取平均值，得到TRE score。

由于有些算法对初始化时给定的bounding box比较敏感，而目前测评用的ground-truth都是人工标注的，因此可能会对某些跟踪算法产生影响。因此为了评估这些跟踪算法是否对初始化敏感，作者通过将ground-truth轻微的平移和尺度的扩大与缩小来产生bounding box。平移的大小为目标物体大小的10%，尺度变化范围为ground-truth的80%到120%，每10%依次增加。最后取这些结果的平均值作为SRE score。

其中一般使用的两个衡量目标跟踪精准度的基本参数是Precision Plot和Success Plot。

**2、VOT数据集（Visual Object Tracking）**

1）简介

VOT数据集是基于每年一次的VOT比赛的，每年都会有新的数据集产生，当然其中一部分图像序列是和OTB重合的，但是总的来说VOT数据集略难于OTB数据集，一般在这两个数据集上跑的效果都好，才算真的好，如果只在一个数据集上效果好，那只能说明这个算法的泛化能力还不够。

2）评价指标

VOT中使用的评价标准EAO和EFO.

EAO（Expect Average Overlap Rate），用于评价性能。EAO的大致思路为：首先将所有序列按照长度分类，令待测tracker在长度为NS的序列上测试，得到每一帧的准确率 :

为一帧的准确率，按重叠百分比定义。长度为NS的序列不止一个，tracker要在这些序列上全跑一遍，这样就得到了该tracker在长度NS序列上的EAO值。之后在其他长度的序列上测EAO，就可以得到一条EAO曲线，横坐标为序列长度，纵坐标为EAO值。

EFO（Equivalent Filter Operations），用于评价速度。

因为性能的比较常常因为计算机性能不同而不够直观，EFO评价标准考虑到了这一点，使用600×600的图像，做30\*30窗口的滤波，来得到机器的性能。然后使用跟踪算法处理每帧图像的评价时间除以以上滤波操作的时间，得到了一个归一化的性能参数，就是EFO，是VOT14提出的标准，一直沿用到现在。

**3.GOT-10K (Generic Object Tracking Benchmark)**

Got-10k是中科院发布的目标跟踪数据集。其中包括超过10000个视频，分为563个类别。目标框超过150万，全部利用手动标注。benchmark可以衡量模型的性能。数据集由动物、人造物体、人物、自然景物和part五大类组成，可细分为563个目标类别。主要评估指标是准确度指标mAO和SR

mAO，AO（average overlap）是指平均重合度，mAO则是考虑到类别平衡，避免多帧序列在平均计算中占据较大权重，故计算公式如下：

C代表种类数，Sc 代表某一种类下的视频序列数。

SR（success rate），成功率。指在一定AO阈值下成功跟踪的准确度，取0.5和0.75两个阈值（即重叠率超过0.5和0.75才算成功）。

* **数据集获取链接**

**OTB：** <http://cvlab.hanyang.ac.kr/tracker_benchmark/datasets.html>

**VOT：** <https://pan.baidu.com/s/1Ck51d7OQ8w8BGcTL9UtopA> （jn0k）

**GOT-1：**

<https://drive.google.com/file/d/1b75MBq7MbDQUc682IoECIekoRim_Ydk1/view?usp=sharing>

**参考文献**

[1] Danelljan; Bhat; Khan; Felsberg. 2019. Atom: Accurate tracking by overlap maximization. In Proceedings of the IEEE Conference on ComputerVision and Pattern Recognition.

[2]Yinda Xu;Zeyu Wang;Zuoxin Li;Ye Yuan, Gang Yu.2020. SiamFC++: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines.

[3]李玺. 2019. 深度学习的目标跟踪算法综述.

[4][Lianghua Huang](https://arxiv.org/search/cs?searchtype=author&query=Huang%2C+L); [Xin Zhao](https://arxiv.org/search/cs?searchtype=author&query=Zhao%2C+X); [Kaiqi Huang](https://arxiv.org/search/cs?searchtype=author&query=Huang%2C+K).2020. GlobalTrack: A Simple and Strong Baseline for Long-term Tracking.

[5]Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8971–8980.

[6]Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

[7]Goutam Bhat, Martin Danelljan, Luc Van Gool, Radu Timofte. 2019. Learning Discriminative Model Prediction for Tracking.

[8]Huang. 2019 Bridging the Gap Between Detection and Tracking: A Unified Approach

[9] Luca Bertinetto , Jack Valmadre, F. Henriques, Andrea Vedaldi. 2016. SiamFC:Fully-Convolutional Siamese Networks for Object Tracking.

[10] Zhu, Z. Wang, Q. Li, B. Wu, W. Yan, J. 2018. Distractor-aware siamese networks for visual object  
tracking.

[11]陈云芳，吴 懿. 2020. 基于孪生网络结构的目标跟踪算法综述.

[12] Comaniciu D，Ramesh V，Meer P.Kernel-based object tracking[J].IEEE Transactions on Pattern Analysis and MachineIntelligence，2003.

[13] Arulampalam M S，Maskell S，Gordon N，et al.A tutorialon particle filters for online nonlinear/non-Gaussian Bayesian tracking[J].IEEE Transactions on Signal Processing，2002.

[14] Fukunaga K，Hostetler L D.The estimation of the gradient of a density function，with applications in patternrecognition[J].IEEE Transactions on Information Theory，1975.

[11] 杨鹏生，吴晓军，张玉梅 . 改进扩展卡尔曼滤波算法的目标跟踪算法[J].计算机工程与应用，2016.

[12] 刘芳，黄光伟，路丽霞，等 . 自适应模板更新的鲁棒目标跟踪算法[J].计算机科学与探索，2019.

[13] Vojir T，Noskova J，Matas J.Robust scale-adaptive mean-shift for tracking[J].Pattern Recognition Letters，2014.

[14] 王忠民，段娜，范琳 . 融合 YOLO 检测与均值漂移的目标跟踪算法[J].计算机工程与应用，2019.

[15] Redmon J，Divvala S K，Girshick R B，et al.You onlylook once：unified，real- time object detection[J].2016.

[16] Bolme D S，Beveridge J R，Draper B A，et al.Visualobject tracking using adaptive correlation filters.

2010.

[17] Henriques J F，Caseiro R，Martins P，et al.Exploiting thecirculant structure of tracking-by-detection with kernels.2012.

[18] Henriques J F，Caseiro R，Martins P，et al.High- speed tracking with kernelized correlation filters[J].IEEE Trans-actions on Pattern Analysis and Machine Intelligence，2015.

[19] Taigman Y，Yang Ming，Ranzato M A，et al.Deep Face：closing the gap to human- level performance in faceverification[C].2014.

[20] Parkhi O M，Vedaldi A，Zisserman A.Deep face recognition.2015.