中国科学院自动化研究所

研究生学位论文中期考核报告

论文题目： 基于孪生卷积网络的实时目标跟踪研究

专　　业： 模式识别与智能系统

研究方向： 视觉目标跟踪

姓　　名： 王　强

学　　号： 201518014628062

培养层次： √博士　□硕士

博士攻读方式： □硕博连读　√直接攻博　□普通招考

导师姓名： 胡卫明

所属部门： **模式识别国家重点实验室**

考核日期： 2019 年 10 月 21 日

目　　录

论文题目

1、研究背景与意义

目标跟踪是计算机视觉领域中近年来备受关注的前沿研究方向。它以摄像机拍摄得到的包含运动目标的序列图像为研究对象,以在视频的连续帧之间创建基于位置、速度、形状、纹理、色彩等有关特征的对应匹配为研究目的,对序列图像中的运动目标进行跟踪。从整个计算机视觉领域来看,它处于中层部分,即既要利用底层的图像处理、运动检测识别信息,同时也为上层的目标行为理解和描述提供了基础,因此具有非常重要的理论研究价值。在这方面,随着 80 年代初光流法的提出,图像序列的跟踪研究开始获得更多的关注。在之后的 30 多年中,众多跟踪算法被提出。同时,在当前的国际权威期刊如 PAMI、IJCV 等以及重要学术会议如 ICCV、CVPR、ECCV 等都将目标跟踪作为重要研究内容列入。

目标跟踪也具有很高的应用价值,它在智能监控、人机交互、虚拟现实、视频分析等方面有着广阔的应用前景。例如,在交通场景中,目标跟踪服务于交通流量控制、车辆异常行为检测、行人行为判定、智能车辆等多个方面；在人机交互环境中,主要解决的是基于计算机视觉的人的手势和表情识别;对于人脸的检测、识别与跟踪还能在远程电视会议中大大降低图像传输的比特率,并使画面及时锁定在讲话人的身上；虚拟现实环境中,通过目标跟踪来确定目标状态,进而为计算机提供合适的场景信息；自动驾驶中，通过目标跟踪捕获车辆周围环境中的障碍物和行人的运动轨迹状态； 另外,目标跟踪的信息也可以应用于视频分析中,例如视频检索。

由于目标跟踪具有这些理论研究价值以及广阔的应用价值,众多研究人员以及公司都投入到这项研究中。同时,随着计算机技术的日益发展,高性能存储和快速计算能力的提升。近三十年来,目标跟踪取得了很多突破,产生了众多的研究成果。但是,目标跟踪领域依旧存在着很多理论和技术问题有待解决,特别是跟踪过程中噪音干扰、运动模糊、光照变化、遮挡等在开放环境中遇到的复杂问题,所以如何能够自适应、实时和鲁棒地跟踪目标一直是广大研究者需要解决的问题,其研究价值高、研究空间依然很大。

1.1 国内外研究现状

Model-free-tracking 是一种基于对初始帧中待跟踪目标只进行单一初始框(bounding box)标注的前提下,广泛适用于对任意目标进行跟踪的跟踪模式。尽管在跟踪特定目标方面,比如人脸、行人、刚性物体等,已经取得了较为显著的成绩,但是在如何跟踪任意指定的目标方面依然具有很大挑战。因为凭借人手工对世间万物标注足够的样本来训练一个特定目标模型,然后用于跟踪是很费时费力的。在这种情况下,Model-free-tracking 这种跟踪模式就有着很大的应用空间。

在 Model-free-tracking 跟踪模式下,我们只对视频序列初始帧中的感兴趣目标区域进行手工标注(一般情况下只使用一个矩形框把目标区域框出来),然后由算法自动完成对剩余视频序列中的感兴趣目标区域的标注,即是完成整个跟踪任务。同时,这个任务的完成必须时刻符合因果性关系(causality),也就是说在任意帧时刻的跟踪结果只能依赖于从初始帧到当前帧的跟踪结果和帧序列,而不能使用任何当前帧之后的帧序列信息。这就使得 MFT (Model-free-tracking)这种跟踪模式有很大的挑战,这是因为: (1)在跟踪一开始,我们只有非常有限的关于待跟踪目标的信息,也就是由初始化 bounding-box 所提供的在初始帧中哪些区域是感兴趣目标区域,哪些不是; (2)同时,这些信息也并非是完全精确无误的,大多都带有一定的含糊性,因为初始化 bounding-box 只能够近似区分感兴趣目标区域和背景区域,而大多数时候初始框所框出的区域中会包含很多背景区域信息;(3)我们感兴趣的目标表观会随着时间由于各种各样的因素而发生剧烈的变化,比如噪音干扰、运动模糊、光照变化、非刚性目标的形变、视角的变化导致的目标旋转、遮挡等。

目前,Model-free-tracking 这种跟踪模式受到了目标跟踪领域内研究者们的广泛关注,并且产生了很多丰硕的研究成果，主要是基于相关滤波的目标跟踪以及基于深度学习的目标跟踪。其中，以全卷积孪生网络为代表的实时深度学习目标跟踪算法引起了大范围的关注，实时的目标跟踪算法也逐步作为目标跟踪的重要约束条件。同时，由于近年来视觉跟踪数据库的不断提出和更新，极大地提升了跟踪领域的发展。吴毅在 CVPR2013 提出的 OTB 视觉跟踪测试集包含 50 段跟踪视频，其中包含多种挑战的场景，其后在 TPAMI2015 上扩展到 100 个视频片段。同时，由伯明翰大学、卢布尔雅那大学、布拉格捷克技术大学、奥地利科技学院联合创办 Visual-Object-Tracking Challenge (VOT)作为国际视觉跟踪领域最权威的测评平台，旨在评测在复杂场景下单目标跟踪的算法性能。其每年举办一次公开的视觉跟踪竞赛为跟踪领域提供良好的评价指标。

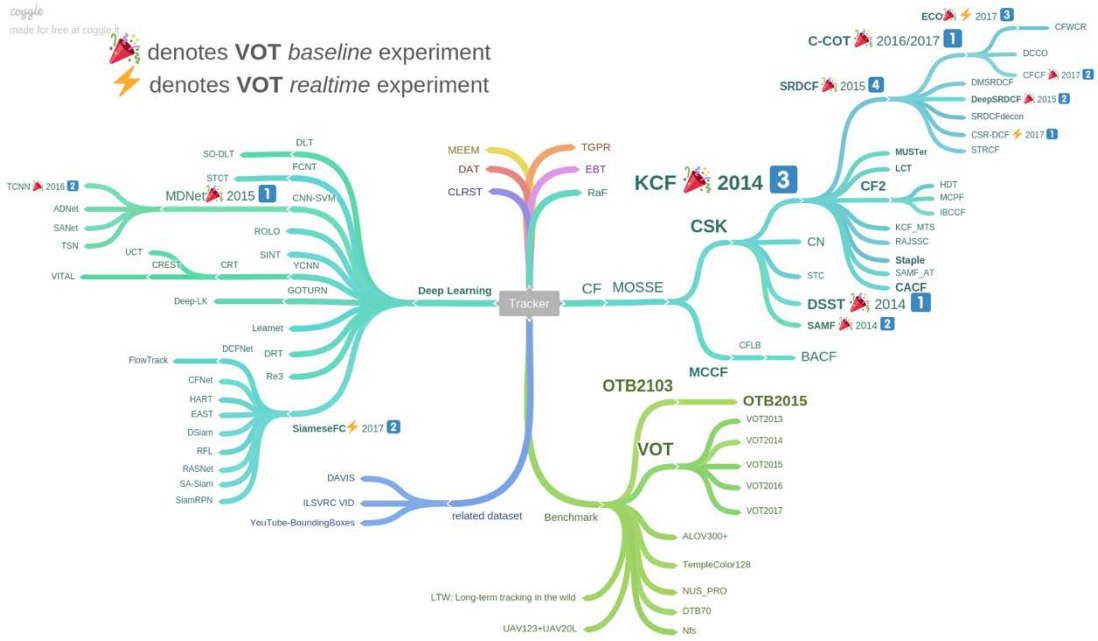


图 1. 目标跟踪的发展结构图

1.1.1 基于相关滤波的目标跟踪

相关是衡量两个信号相似值的度量,如果两个信号越相似,那么其相关值就越高。而在视觉跟踪里,就是需要设计一个滤波模板,使得当它作用在跟踪目标上时,得到的响应最大。相关滤波方法最大的优势在于其速度之快,是任何其他跟踪方法都无法比拟的,如 MOSSE[1],其速度可达 669 帧每秒,把跟踪算法从实时级别提升到了高速级别;而且其跟踪准确率高,在 CVPR2013 发表的标准库上,带核函数的 CSK 方法[2]可以得到 73%左右的准确率。CSK 跟踪器的最大亮点就是提出了利用循环移位的方法进行稠密采样并结合 FFT 快速地进行分类器的训练。稠密采样的采样方式能提取目标的所有信息,这对目标的跟踪至关重要。虽然 CSK 的速度很快,但是 CSK 只是简单的使用了灰度特征,对目标的外观描述能力显然不足。对此作者改进了 CSK 提出了 KCF[3],从原来的单通道灰度特征换成了多通道 HOG 特征[4]。KCF 算法通过核函数对多通道的 HOG 特征进行了融合,使得训练所得的分类器对待检测目标的解释力更强。KCF 跟踪器存在对光照变化,遮挡,非刚性形变,运动模糊,背景杂乱和旋转等情况均能跟踪良好,但对尺度变化,快速运动,刚性形变等情况的视频跟踪效果不佳。和 KCF 同一时期的还有个 CN[5],它是在 2014 年 CVPR 上引起剧烈反响的颜色特征方法,也是 CSK 的多通道颜色特征改进算法。随后,DSST[6]、SAMF[7]算法通过引入尺度滤波器以及在搜索目标过程

中引入尺度变化来提高跟踪器对于目标尺度变化的估算准确率。基于相关滤波的跟踪器受到边缘效应的影响,往往对目标的快速运动不鲁棒。SRDCF[8]算法通过加入空域正则化来解决边缘效应的问题,惩罚边界区域的滤波器系数,忽略所有移位样本的边界部分像素,让边界附近滤波器系数接近为零。基于灰度特征的 CFLM 算法[9]和基于 HOG 特征的 BACF[10]算法也一定程度解决了边缘效应问题,主要思路是采用较大尺寸检测图像块和较小尺寸滤波器来提高真实样本的比例,即从大尺寸样本的循环移位样本集中用掩膜矩阵裁切出真实的小尺寸样本,然后与小尺寸滤波器进行循环相关操作。CCOT 算法[11]是 VOT2016 的第一名,综合了 SRDCF[8]的空域正则化和 SRDCFdecon[12]的自适应学习样本权重,还将 DeepSRDCF[13]的单层卷积的深度特征扩展为多层卷积的深度特征(VGG 的第 1 和 5 层),最重要的是为了应对不同卷积层分辨率不同的问题,提出了连续空间域插值转换操作,在训练之前通过频域隐式插值将特征图插值到连续空域,方便集成多分辨率特征图,并且保持定位的高精度,目标函数通过共轭梯度下降方法迭代优化。CFNet[15]把相关滤波和深度特征结合，实现端到端的训练。

1.1.2 基于深度学习的目标跟踪

不同于检测、识别等视觉领域中深度学习一统天下的趋势,深度学习在目标跟踪领域的应用并非一帆风顺。其主要问题在于训练数据的缺失:深度模型的魔力之一来自于对大量标注训练数据的有效学习,而目标跟踪仅仅提供第一帧的 bounding-box 作为训练数据。这种情况下,在跟踪开始针对当前目标从头训练一个深度模型困难重重。为了解决这个问题，目前基于深度学习的目标跟踪算法主要采用了几种思路。

1.1.3 利用辅助图片数据预训练深度模型,在线跟踪时微调

在目标跟踪的训练数据非常有限的情况下,使用辅助的非跟踪训练数据进行预训练,获取对物体特征的通用表示(general representation),在实际跟踪时,通过利用当前跟踪目标的有限样本信息对预训练模型微调(fine-tune),使模型对当前跟踪目标有更强的分类性能,这种迁移学习的思路极大的减少了对跟踪目标训练样本的需求,也提高了跟踪算法的性能。这个方面代表性的成果有 DLT[16]和 SO-DLT[17],都出自香港科技大学王乃岩博士。

1.1.4 利用现有大规模分类数据集预训练的 CNN 分类网络提取特征

2015 年以来,在目标跟踪领域应用深度学习兴起了一股新的潮流。即直接使用 ImageNet 这样的大规模分类数据库上训练出的 CNN 网络如 VGG-Net 获得目标的特征表示,之后再用表观模型进行分类获得跟踪结果。这种做法既避开了跟踪时直接训练大规模CNN样本不足的困境,也充分利用了深度特征强大的表征能力。这样的工作在 ICML15,ICCV15,CVPR16 均有出现。

作为应用 CNN 特征于物体跟踪的代表作品,FCNT[18]的亮点之一在于对 ImageNet 上预训练得到的 CNN 特征在目标跟踪任务上的性能做了深入的分析,并根据分析结果设计了后续的网络结构。FCNT 主要对 VGG-16 的 Conv4-3 和 Conv5-3 层输出的特征图谱做了分析,并得出以下结论:(1)CNN的特征图谱可以用来做跟踪目标的定位。(2)CNN 的许多特征图谱存在噪声或者和物体跟踪区分目标和背景的任务关联较小。(3)CNN 不同层的特征特点不一。高层(Conv5-3)特征擅长区分不同类别的物体,对目标的形变和遮挡非常鲁棒,但是对类内物体的区分能力非常差。低层(Conv4-3)特征更关注目标的局部细节,可以用来区分背景中相似的干扰项,但是对目标的剧烈形变非常不鲁棒。

依据以上分析,FCNT 最终的框架结构:(1)对于 Conv4-3 和 Conv5-3 特征分别构建特征选择网络 sel-CNN(1 层 dropout 加 1 层卷积),选出和当前跟踪目标最相关的特征图谱。(2)对筛选出的 Conv5-3 和 Conv4-3 特征分别构建捕捉类别信息的 GNet 和区分干扰项(背景相似物体)的 SNet(都是两层卷积结构)。(3)在第一帧中使用给出的标注信息生成热度图回归训练 sel-CNN,GNet 和 SNet。(4)对于每一帧,以上一帧预测结果为中心提取出感兴趣区域,之后分别输入 GNet 和 SNet,得到两个预测的热图,并根据是否有干扰项决定使用哪个热图生成最终的跟踪结果。总而言之,FCNT根据对 CNN不同层特征的分析,构建特征筛选网络和两个互补的定位预测网络。达到有效抑制干扰项防止跟踪器漂移,同时对目标本身的形变更加鲁棒的效果,也是集成思路的又一成功实现。在 CVPR2013 提出的 OTB50 数据集上准确度绘图达到了 0.856,成功率绘图达到了 0.599,准确度绘图有较大提高。实际测试中 FCNT 的对遮挡的表现不是很鲁棒,现有的更新策略还有提高空间。

1.1.5 利用跟踪序列预训练, 在线跟踪时微调

意识到图像分类任务和跟踪之间存在巨大差别,MDNet[19]提出直接用跟踪视频预训练 CNN 获得目标表示的方法。但是序列训练也存在问题,即不同跟踪序列跟踪目标完全不一样,某类物体在一个序列中是跟踪目标,在另外一个序列中可能只是背景。不同序列中目标本身的表观和运动模式、环境中光照、遮挡等情形相

差甚大。这种情况下,想要用同一个 CNN 完成所有训练序列中前景和背景区分的任务,困难重重。最终 MDNet 提出 Multi-Domain 的网络训练思路。该网络分为共享层和 domain-specific 层两部分。即将每个训练序列当成一个单独的 domain,每个 domain 都有一个针对它的二分类层(fc6),用于区分当前序列的前景和背景,而网络之前的所有层都是序列共享的。这样共享层达到了学习跟踪序列中目标通用的特征表达的目的,而 domain-specific 层又解决了不同训练序列分类目标不一致的问题。MDNet 的训练数据也很有讲究,即测试 OTB100 数据集时,利用 VOT2013-2015 的不重合的 58 个序列来做预训练。测试 VOT2014 数据集时,利用 OTB100 上不重合的 89 个序列做预训练。这种交替利用的思路也是第一次在跟踪论文中出现。MDNet 从检测任务中借鉴了不少行之有效的策略,如难例挖掘,边框回归等。尤其是难例挖掘通过重点关注背景中的难点样本(如相似物体等)显著减轻了跟踪器漂移的问题。这些策略也帮助 MDNet 在 OTB100 数据集上准确度绘图从一开始的0.825提升到0.908,成功率绘图从一开始的0.589提升到0.673。

但是也可以发现 MDNet 的总体思路和 RCNN 比较类似,需要前向传递上百个候选边框,虽然网络结构较小,速度仍较慢。且边框回归也需要单独训练,因此 MDNet 还有进一步提升的空间。

1.1.6 基于全卷积孪生神经网络的目标跟踪算法

为了提升深度学习的速度，Bertinetto 在 ECCV2016 workshop 中提出 SiamFC 算法[20]。该算法离线过程中通过大量样本使用全卷积孪生神经网络进行度量学习。在线跟踪过程中不更新网络参数，使得算法可以在 GPU 上运行超过 80FPS，该算法在 VOT2017 竞赛中取得 realtime 的冠军。随后 Huang 基于此架构设计自适应级联孪生网络 EAST[21]，通过对响应状态进行分析，动态调整网络的深度，使得算法在 CPU 上可以达到实时运行的速度。Li 在 CVPR2018 提出使用基于全卷积孪生区域提议网络 SiamRPN[22]，通过多锚点的回归来得到长宽比自适应的跟踪算法。该算法极大地提升了网络的精度，获得 VOT2018 竞赛的实时跟踪冠军。

2、学位论文撰写提纲

学位论文预计分为七个章节。
第一章为绪论，介绍研究背景和意义，研究内容和论文主要贡献，以及论文结构安排。
第二章为研究现状，介绍单目标跟踪领域的已有方法及各自优缺点，以及我自己的研究工作与它们的联系。
第三章为基于可微分的相关滤波框架的目标跟踪研究，介绍我的第一个研究工作，包括引言、算法推导与实验分析。
第四章为基于注意力机制的孪生网络目标跟踪研究，介绍我的第二个研究工作。
第五章为基于编解码架构的相关滤波训练的自适应跟踪研究，介绍我的第三个研究工作。
第六章为介绍目标跟踪和视频分割的统一框架，介绍我的第四个研究工作。
第七章为总结与展望，总结我的研究工作的贡献和不足，展望未来可以继续研究的内容。

3、学位论文主要创新成果

从开始研究工作至今,已完成了四项目标跟踪方面的研究:1)基于端到端学习相关滤波框架的目标跟踪。通过对相关滤波前后向传播的研究推导，将相关滤波算法的特征进行调整，得到针对于相关滤波跟踪任务的特征网络，使相关滤波的精度得到提升。最终,形成一个完整的基于相关滤波的跟踪框架,在跟踪测评集OTB2013、OTB2015，以及 VOT2016 上取得了良好的成绩，相关论文现已发表于ICPR2018，其 Arxiv 版本 Google Scholar 引用超过 90 次。2)基于残差注意力机制的孪生神经网络目标跟踪。为了解决孪生网络目标跟踪的匹配过程中的空间权重的分布估计问题,提出残差注意力机制，通过通用注意力和残差注意力的结合提升网络的判别能力。该项工作已发表在 CVPR2018 上，Google Scholar 引用超过 100 次。3)基于编解码网络的孪生神经网络目标跟踪。在判别跟踪学习过程中引入编解码结构,将生成式自监督损失引入联合学习过程,增加网络的泛化性能。同时，通过将底层的相关滤波跟踪器和高层的孪生网络跟踪器进行有效融合，提升网络的鲁棒性。该项工作已发表在 IJCAI2018 上。4)基于多分枝预测的孪生神经网络目标跟踪。在目标空间位置响应图预测的基础上增加分割分支，使得网络

可以得到目标的高精度表述。同时，通过分割分支的引入，形成目标跟踪和视觉目标分割一体化的架构。该项工作已发表在 CVPR2019 上。下面分别介绍这四项研究工作。

3.1 基于端到端学习相关滤波框架的目标跟踪

判别相关滤波（DCF）方法目前在跟踪领域占据主导地位。但是其特征提取部分通常采用手工设计特征（例如 HOG，CN）或在 ImageNet 上训练的卷积神经网络特征。本文提出端到端学习的轻量化网络架构来学习适应于相关滤波的卷积神经网络。我们将判别相关滤波作为一个特殊的网络层，通过推导其前后向传播过程加入到孪生神经网络的架构当中，通过对网络输出与预定义的概率分布直接的差异来优化网络参数。由于前后向推导都是在频域中实现，保留了原有判别相关滤波的速度，这使得跟踪算法可以在测试过程中超过 60 帧每秒（FPS）。同时由于自适应的特征的提出，使得算法的精度大大提升。

在传统的判别相关滤波算法框架中，首先在目标区域截取目标图像，并通过卷积网络抽取特征$\varphi(x)\in R^{M\times N\times D}$，通过与理想的回归响应$y\in R^{M\times N}$直接构建判别回归学习。通过最小化回归损失，可以得到利用训练样本学习的滤波器$w\in R^{M\times N\times D}$：

$$\varepsilon=\sum_{l=1}^{D}\left\|w^{l}*\varphi^{l}(x)-y\right\|^{2}+\lambda\sum_{l=1}^{D}\left\|w^{l}\right\|^{2} \tag{1}$$

该问题存在一个高效的频域闭式解，其中$\hat{w}$表示w的傅里叶变换，y^{*}表示y的共轭，$\odot$表示点乘：

$$\hat{w}^{l}=\frac{\hat{\varphi}^{l}(x)\odot\hat{y}^{*}}{\sum_{l=1}^{D}\hat{\varphi}^{l}(x)\odot(\hat{\varphi}^{l}(x))^{*}+\lambda} \tag{2}$$

对于在线跟踪过程，截取同样大小的搜索区域，并通过特征提取网络提取其特征表述$\varphi(z)\in R^{M\times N\times D}$。通过在频域的操作快速得到目标的响应，将最大响应位置设置为新的目标位置，依次迭代进行跟踪。

$$g=F^{-1}(\sum_{l=1}^{D}\hat{\varphi}^{l}(z)\odot(\hat{w}^{l})^{*}) \tag{3}$$

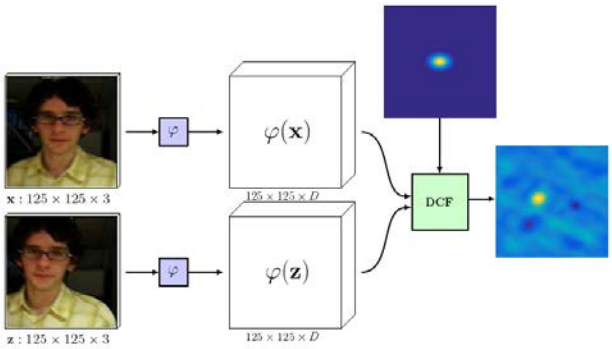


图 2. 判别相关滤波网络框架图

对于传统的判别相关滤波算法，只有在线的超参数可以进行启发式的搜索，而本文提出的方法可以同时调节特征网络参数预计在线更新参数。该方法的结构框架如图 2 所示，网络架构通过级联特征提取层和判别相关滤波层来获得目标的位置。在离线过程中，通过优化目标响应 g 和理想响应 $\tilde{g}$ 之间的差值来获得训练的目标函数：

$$L(\theta)=\left\|g-\tilde{g}\right\|^2+\gamma\left\|\theta\right\|^2 \tag{4}$$

其前项过程与传统的相关滤波保持一致；本方法通过在反向传播过程中的推导实现网络的端到端学习。首先，分析离散傅里叶变换的基本推导形式：

$$\hat{g}=F(g),\frac{\partial L}{\partial \hat{g}^*}=F(\frac{\partial L}{\partial g}),\frac{\partial L}{\partial g}=F^{-1}(\frac{\partial L}{\partial \hat{g}^*}) \tag{5}$$

由于前项过程中存在点乘操作，损失函数可以根据每个位置进行反向传播：

$$\frac{\partial L}{\partial \hat{g}^*_{uv}}=(F(\frac{\partial L}{\partial g}))_{uv} \tag{6}$$

对于检测分支的反向传播，可以通过下式得到：

$$\frac{\partial L}{\partial(\hat{\varphi}^l_{uv}(z))^*}=\frac{\partial L}{\partial \hat{g}^*_{uv}}\frac{\partial \hat{g}^*_{uv}}{\partial(\hat{\varphi}^l_{uv}(z))^*}=\frac{\partial L}{\partial \hat{g}^*_{uv}}(\hat{w}^l_{uv}) \tag{7}$$

对于学习分支的反向传播，将 $\hat{\varphi}^l_{uv}(x)$ 和 $(\hat{\varphi}^l_{uv}(x))*$ 当作独立变量分别求导：

$$\frac{\partial L}{\partial \hat{\varphi}^l_{uv}(x)}=\frac{\partial L}{\partial \hat{g}^*_{uv}}\frac{(\hat{\varphi}^l_{uv}(z))^*\hat{y}^*_{uv}-\hat{g}^*_{uv}(\hat{\varphi}^l_{uv}(x))^*}{\sum_{l=1}^{D}\hat{\varphi}^l_{uv}(x)(\hat{\varphi}^l_{uv}(x))^*+\lambda} \tag{8}$$

$$\frac{\partial L}{\partial(\hat{\varphi}^l_{uv}(x))*}=\frac{\partial L}{\partial \hat{g}^*_{uv}}\frac{-\hat{g}^*_{uv}\hat{\varphi}^l_{uv}(x)}{\sum_{l=1}^{D}\hat{\varphi}^l_{uv}(x)(\hat{\varphi}^l_{uv}(x))^*+\lambda} \tag{9}$$

通过将判别相关滤波层的反向传播推导，可以将目标函数的梯度传回到特征层，这使得整体优化特征网络和判别相关滤波的端到端的优化可以实现。

3.1.1 实验结果

本文提出的 SPCNet 算法在跟踪三大标准库 OTB2013、OTB2015、VOT2015 上进行了测试体现了其高效性。实验结果如下所示，表 1 显示了 SPCNet 算法相对于传统 DCF 的精度提升，相对于使用 VGG 网络的算法，算法的各项精度均有提升。图 3 展示的算法的多种变种形式，可以根据应用场景调节得到 60~180FPS 的不同算法。图 4 展示的是在 OTB2013、OTB2015 上的成功率曲线，跟踪器按照 AUC（area under the curve）值进行排名。图 5 展示的是在 VOT2015 库上的综合性能 EAO。

Trackers	OTB-2013		OTB-2015		FPS
	OP	DP	OP	DP	
DCF [7]	61.6	72.8	54.8	68.9	**292**
DCFNet1s	67.7	79.1	63.7	76.8	*187*
SAMF [30]	67.7	78.5	64.0	74.3	12
DSST [11]	67.1	74.7	60.9	68.9	46
DCF+VGG	62.1	66.1	61.7	66.9	88
DCF+SiamFC	66.8	74.2	64.0	68.0	77
DCFNet	*78.5*	*86.7*	*72.8*	*79.4*	109
SPCNet	**84.3**	**88.1**	**77.6**	**82.9**	67

表 1. OTB2013 和 OTB2015 库上与传统相关滤波算法对比

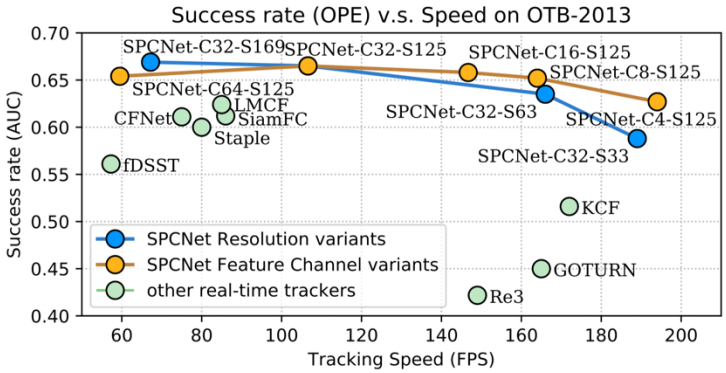
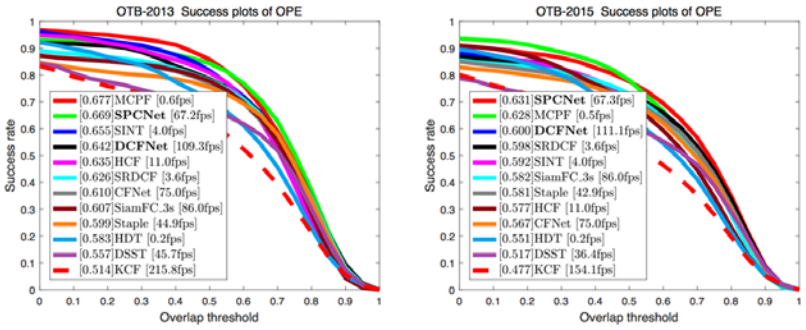


图 3. SPCNet 的各种变种与其他实时算法的速度精度对比



(a) OPE on OTB-2013　　(b) OPE on OTB-2015

图 4. SPCNet 在 OTB2013、OTB2015 成功率曲线

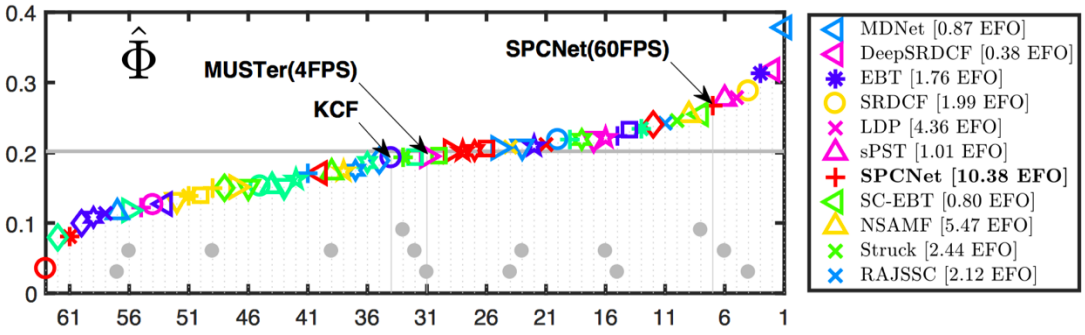


图 5. SPCNet 在 VOT2015 的期望重叠率曲线

3.2 基于残差注意力的孪生神经网络目标跟踪

离线训练的孪生网络目标跟踪由于其速度和精度的均衡表现，近来受到了极大的关注。本文提出残差注意力孪生网络（RASNet）来进行在线目标跟踪。通过使用孪生网络的方式重新形式化了相关滤波跟踪算法，通过引入不同的注意力机制来自适应调整模型。利用离线训练得到通用注意力和目标自适应的残差注意力，以及通道感知的注意力。通过注意力机制的学习，该算法不仅减轻了神经网络的过拟合问题，同时通过将判别学习和特征学习相分离提升了网络的判别能力。在 OTB2015 和 VOT2017 测试集上的大量实验证明本文所提出的算法可以取得当前最好的跟踪精度的同时，保持超过 80FPS 的跟踪速度。

为了得到高效的跟踪算法，我们提出残差注意力机制网络。图 6 是本文提出的算法结构图。

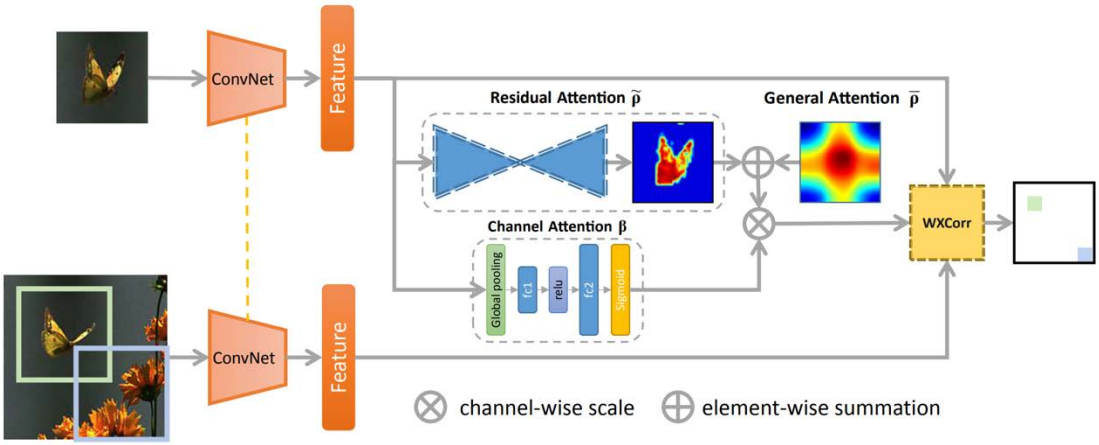


图 6. 残差注意力孪生神经网络结构图

3.2.1 孪生网络简介

目标跟踪问题可以被形式化为一个回归问题：

$$\min_{\mathbf{w}} \|\mathbf{A}\mathbf{w}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2 \tag{10}$$

其中$\mathbf{A}$是训练特征集合，$\mathbf{y}$是相应的标签结果，$\|\bullet\|_2$表示L2模长。问题的解可表述为：

$$\mathbf{w}=(\mathbf{A}^T\mathbf{A}+\lambda\mathbf{I})^{-1}\mathbf{A}^T\mathbf{y} \tag{11}$$

由于上式中存在矩阵的逆的计算，使得问题的计算复杂度较高。通常会在对偶空间进行求解。

$$\mathbf{w}=\mathbf{A}^T\boldsymbol{\alpha} \tag{12}$$

我们可以看到这种对偶形式实际上是将特征提取和判别学习相分离，$\boldsymbol{\alpha}$用来反映判别部分。对于孪生网络而言，通过离线进行度量学习$f(\mathbf{x},\mathbf{z})$，最大化目标样例$\mathbf{z}$与搜索空间中的负样本$\mathbf{x}$的距离，最小化目标和搜索空间中的正样本的距离。在更大的搜索空间中通过卷积网络进行共享计算，得到每个位置的响应图。

$$f(\mathbf{x},\mathbf{z})=\phi(\mathbf{x})*\phi(\mathbf{z})+b \tag{13}$$

由上式可以看出，对于孪生网络跟踪算法需要利用同一个网络同时进行特征提取和判别学习，这使得原有的孪生网络算法更容易过拟合。

为了克服上述孪生网络跟踪算法的缺点，相关滤波网络提出使用循环样本在线学习。但是该算法存在边缘效应，无法适应低分辨率的特征表述。

3.2.2 加权重的交叉相关

为了克服孪生网络的不足，本文提出加权重的相关滤波算法。为了调节不同空间位置对最终相似性度量，该算法扩展了原有的交叉相关算法。交叉相关可以具体表述为在给定目标特征$\phi(z)\in R^{m\times n\times d}$，搜索区域特征表述为$\phi(x)\in R^{p\times q\times d}$，对于目标响应$f\in R^{p'\times q'}$，其中$p\geq m,q\geq n,p'=p-m+1,q'=q'-n+1$。

$$f_{p''q''}=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{c=0}^{d-1}\phi_{i,j,c}(\mathbf{z})\phi_{p''+i,q''+j,c}(\mathbf{x})+b \tag{14}$$

而实际跟踪过程中，每个$\phi(\mathbf{z})$中每个位置对于最终相似度的度量的贡献并不一致，因而本文提出加权中的相关滤波操作来区别每个位置的重要性。

$$f_{p''q''}=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{c=0}^{d-1}\gamma_{i,j,c}\phi_{i,j,c}(\mathbf{z})\phi_{p''+i,q''+j,c}(\mathbf{x})+b \tag{15}$$

$$f_{p''q''}=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\sum_{c=0}^{d-1}\rho_{i,j}\beta_c\phi_{i,j,c}(\mathbf{z})\phi_{p''+i,q''+j,c}(\mathbf{x})+b \tag{16}$$

$$f(\mathbf{x},\mathbf{z})=\phi(\mathbf{x})*(\gamma\odot\phi(\mathbf{z}))+b \tag{17}$$

而γ的作用可以认为是目标特征的注意力，本文提出使用神经网络来学习这个注意力机制。启发式地，对于视觉跟踪目标而言，中心点的注意力应该会比边缘的注意力更为重要。带权重的交叉相关可以通知编码空间位置的重要性以及不同特

征层的重要性。然而通过神经网络直接得到γ会引入过多的计算量，因而将整体的注意力分解为空间对偶注意力ρ以及通道注意力β。这使得需要估计的参数量从$m \cdot n \cdot d$减少到$m \cdot n + d$，降低训练的难度。

3.2.3 空间对偶注意力

空间对偶注意力ρ可以通过神经网络学习得到，一种注意力机制是约束所有的数据来共享一个通用的注意力。我们首先通过$m \cdot n$个初始值为 1 的变量来训练得到通用注意力$\bar{\rho}$。然后利用残差网络来学习使用与每个目标的残差注意力$\tilde{\rho}$：

$$\rho = \bar{\rho} + \tilde{\rho} \tag{18}$$

通用注意力$\bar{\rho}$编码了来自所有训练数据的通用信息，而残差部分$\tilde{\rho}$描述了不同目标各自的判别信息。对于不同目标通过残差部分进行自适应的调整。

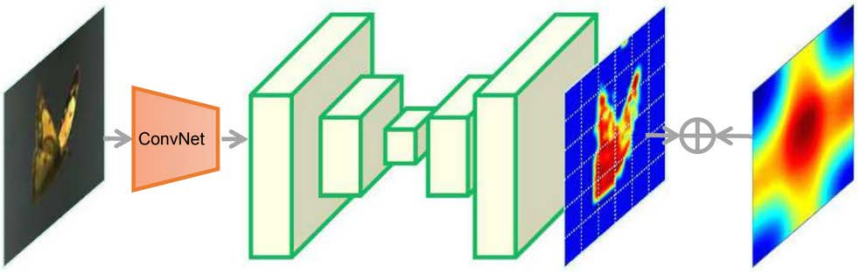
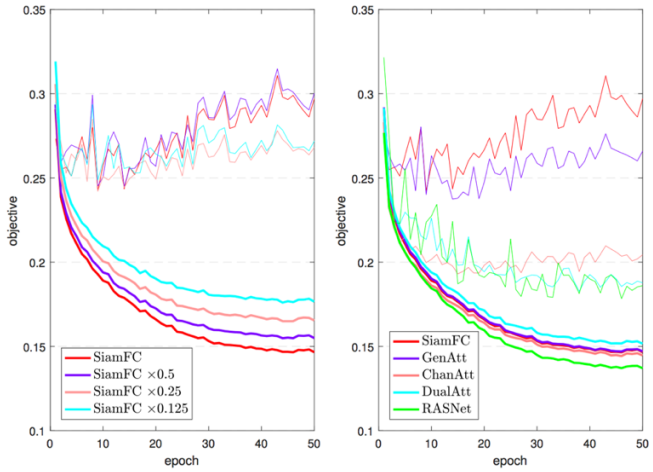


图 7. 空间对偶注意力的结构图

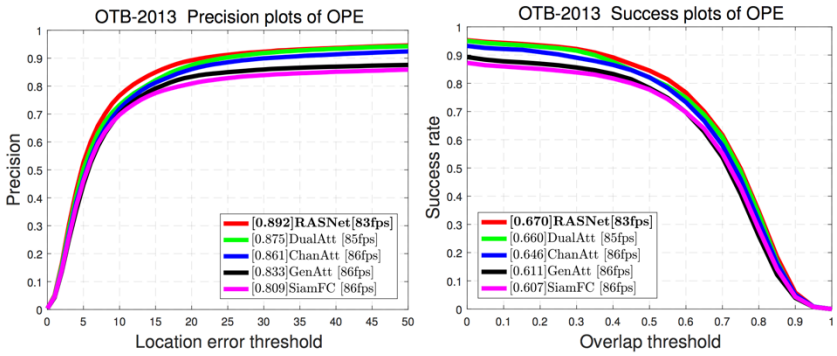
3.2.4 实验结果

首先我们分析了算法的训练过程，可以看到，相对于原始的孪生全卷积网络（SiamFC），残差注意力机制跟踪算法的训练过程收敛性更好。在验证集上的表现表明该算法的泛化性能更强。



(a) SiamFC [2] family (b) RASNet family

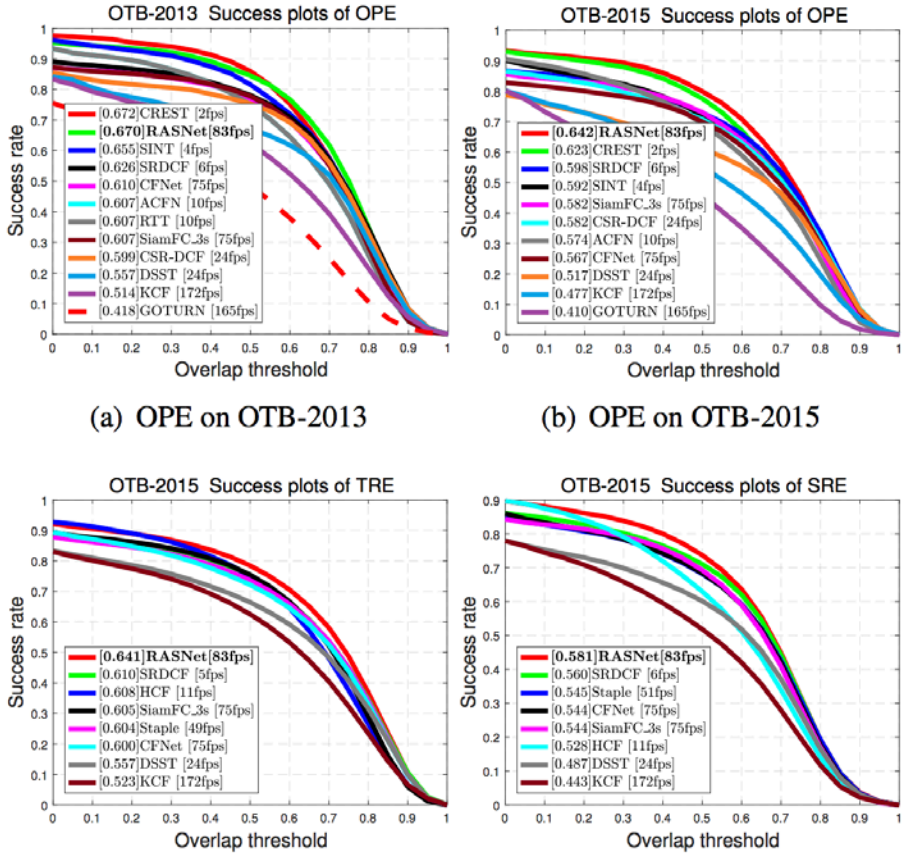
图 8. SiamFC 算法和 RASNet 算法的训练曲线



(a) Precision plot (b) Success plot

图 9. 多种注意力机制在 OTB2013 上的精度曲线

在跟踪标准库 OTB2013 和 OTB2015 上测试了 RASNet 算法。结果如图 10 所示，可以看到本文提出的算法在 OTB2013 和 OTB2015 上均取得了较好的性能，同时，算法的速度要远快于当其他高精度算法。



(a) OPE on OTB-2013 (b) OPE on OTB-2015

(c) TRE on OTB-2015 (d) SRE on OTB-2015

图 10. RASNet 在 OTB2013 和 OTB2015 上的成功率曲线

同时在VOT2015和VOT2017测评库上，本文提出的算法可以取得较为理想的精度。更为重要的是，在VOT2017测评库上，该算法可以取得当前最好的实时跟踪精度。

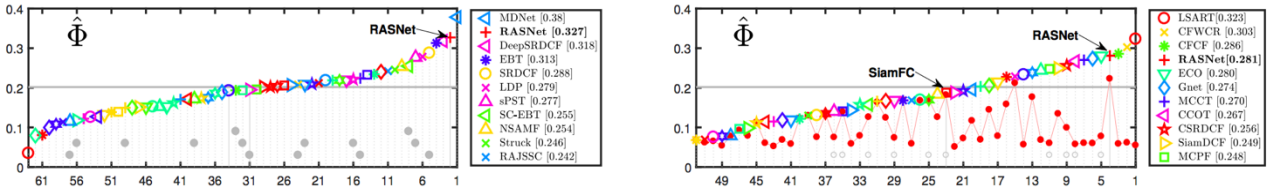


图 11. RASNet 在 VOT2015 和 VOT2017 上的期望覆盖率

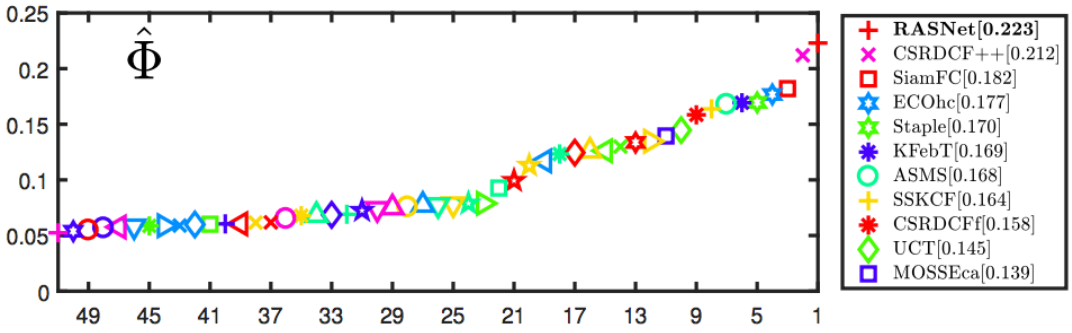


图 12. RASNet 在 VOT2017 实时设置中的期望覆盖率

3.3 基于编解码网络的孪生神经网络目标跟踪

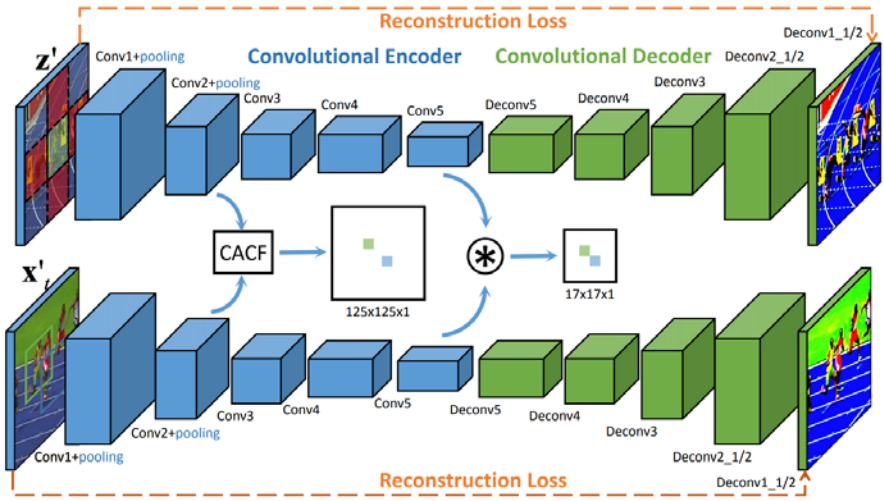


图 13. 编解码相关滤波的框架图

在前两项工作的基础上，我们提出了端到端的编解码网络。该网络同时利用目标的底层特征和高层特征进行跟踪，增加高层的细粒度描述能力。本文所构建的孪生网络框架可以同时学习底层的细节表述，同时利用高层的语义描述进行互补增强。此外，我们设计了一个全卷积的编解码网络用来通过语义投影重建原始

目标图像，这样的设计使得语义投影保留了所有的结构信息。对于底层的相关滤波学习，我们引入了上下文环境约束，增加了相关滤波网络的判别性能。通过基于高层语义特征的交叉相关和底层自适应相关滤波的融合，使得算法的鲁棒性和精度都有较大提升。

3.3.1 通用语义嵌入空间学习

不同于近期的深度学习跟踪算法只关注与判别式学习，本文提出通过引入额外的重构误差来增加高层表述网络的泛化性能。由于图像的重构是一个自监督的任务，相比于判别式学习对于训练数据的分布较为鲁棒。这使得算法的语义嵌入层更加鲁棒，增加了整体跟踪的鲁棒性。更为重要的是，重建约束使得语义嵌入空间保留了所有的结构信息，增加了跟踪算法的精度。

本文提出的通用语义嵌入学习是基于编解码网络架构（encoder-decoder）。其中编码器网络 ϕ： $R^{M\times N\times 3}\rightarrow R^{P\times Q\times D}$ 由 5 个卷基层以及两个最大池化层（Max Pooling）组成，将原始图像表述投影到语义嵌入空间。解码器 ψ：$R^{P\times Q\times D}\rightarrow R^{M\times N\times 3}$ 将低分辨率的语义描述投影到图像空间，并且还原到原始图像的分辨率。解码器部分由 7 个解卷积网络实现。因而通用语义嵌入框架的学习通过最小化重构损失 L_{recon} 和跟踪损失 L_{high} 进行联合学习。

$$L_{sel}=L_{high}+L_{recon} \tag{19}$$

$$L_{recon}=\left\|\psi(\phi(z';\theta_e))-z'\right\|_2+\left\|\psi(\phi(x';\theta_e))-x'\right\|_2 \tag{20}$$

对于高层语义跟踪架构，通过采用内积来度量目标图像和搜索图像的相似度：

$$f_{u,v}=\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}\left\langle\phi_{i,j}(z';\theta_e),\phi_{i,j}(x';\theta_e)\right\rangle \tag{21}$$

高层语义跟踪的损失函数采用 Logistic 损失：

$$L_{high}=\frac{1}{|D|}\sum_{(u,v)\in D}\log(1+\exp(-y_{u,v}f_{u,v})) \tag{22}$$

3.3.2 上下文空间感知的自适应相关滤波跟踪

尽管重建约束使得语义空间具有底层的结构信息，但直接在底层高分辨率特征层进行相关分析依然十分重要。我们提出引入全局的上下文约束到相关分析来抑制干扰物。我们同样构建一个端到端学习的相关滤波层来在线进行更新调整，该方法在频域进行高效运算，保留了算法实时的性能。

我们首先对通用的相关滤波（CF）算法进行总结。相关滤波通过在目标特征

进行稠密采样，将所有的平移样本连接得到特征矩阵$\mathbf{Z}_0$。通常特征提取部分采用手工特征或者预训练好的卷积神经网络。

$$\min_{\mathbf{w}} \|\mathbf{Z}_0\mathbf{w}-\mathbf{y}\|_2^2+\lambda\|\mathbf{w}\|_2^2 \tag{23}$$

这种循环移位结构可以通过频域得到高效的闭式解。通过引入全局上下文进行正则来增加模型的判别能力。在目标图像附近我们采样 k 个上下文样本z_i'，这些上下文样本可以看作多种干扰物和不同的背景。我们约束这些上下文背景的响应为 0：

$$\min_{\mathbf{w}} \|\mathbf{Z}_0\mathbf{w}-\mathbf{y}\|_2^2+\lambda_1\|\mathbf{w}\|_2^2+\lambda_2\sum_{i=1}^{k}\|\mathbf{Z}_i\mathbf{w}\|_2^2 \tag{24}$$

该优化目标在频域的闭式解为：

$$\hat{\mathbf{w}}=\frac{\hat{z}_0^*\odot\hat{y}}{\hat{z}_0^*\odot\hat{z}_0+\lambda_1+\lambda_2\sum_{i=1}^{k}\hat{z}_i^*\odot\hat{z}_i} \tag{25}$$

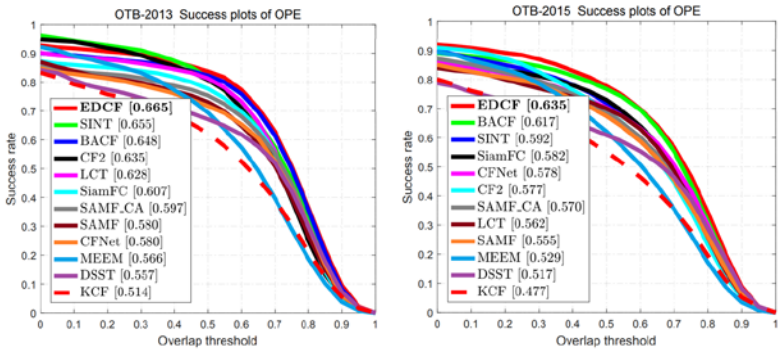
3.3.3 实验结果

首先在 OTB2013 和 OTB2015 以及 VOT2015 上进行对照实验分析，通过比较原始的孪生网络（SiamFC）和我们的编解码孪生网络（EDSiam）。通过引入编解码的结构，使得网络的判别能力得到提升，在距离精度指标中提升超过 3%。带有上下文的相关滤波算法 CACFNet 也极大地提升了网络的判别能力，提升了相关滤波的精度，在重叠率精度指标中提升超过 12%。最终两者的融合算法(EDCF)在所有评价指标上均取得最好的跟踪结果，同时保持了实时跟踪的速度。

Trackers	OTB-2013		OTB-2015		VOT15	FPS
	OP	DP	OP	DP	EAO	
SiamFC	77.8	80.9	73.0	77.0	0.289	*86*
EDSiam	79.0	83.9	75.4	80.7	*0.293*	*86*
CFNet	71.7	76.1	70.3	76.0	0.217	75
CACF	75.4	80.3	68.9	79.1	0.199	13
CACFNet	83.8	87.6	77.7	82.7	0.271	**109**
CACFNet+	*83.9*	*88.3*	*78.0*	*83.1*	0.277	**109**
EDCF	**84.2**	**88.5**	**78.5**	**83.6**	**0.315**	65

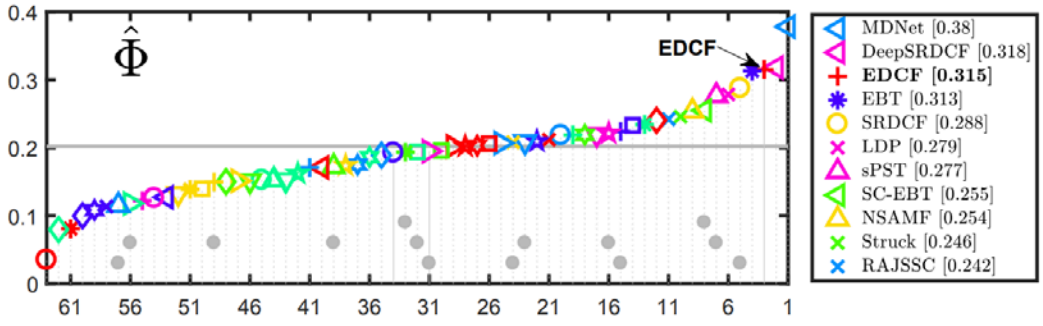
表 2. EDCF 的变种算法与和传统相关滤波算法的速度精度对比

在跟踪标准库 OTB2013 和 OTB2015 上测试了 EDCF 算法。结果如图所示。可以看到该算法在 OTB2013 和 OTB2015 上均取得了较好的性能，同时，本文提出的算法的速度要远快于当其他高精度算法。

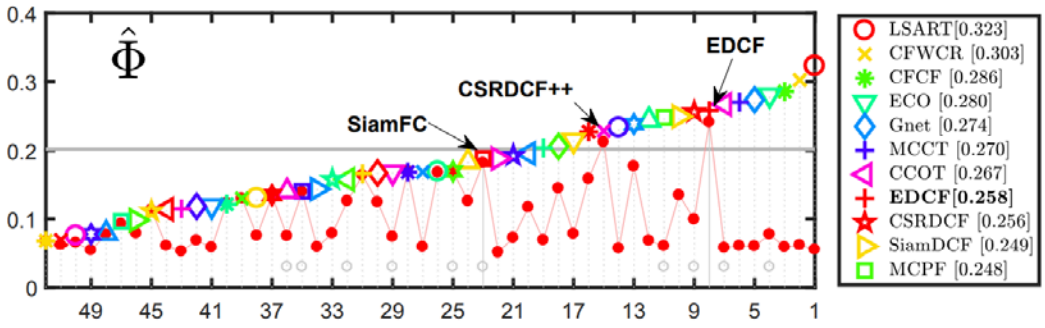


(a) OPE on OTB2013　　(b) OPE on OTB2015

图 14. EDCF 在 OTB2013 和 OTB2015 上的成功率曲线



(a) EAO on VOT2015



(b) EAO on VOT2017

图 15. EDCF 在 VOT2015 和 VOT2017 上的期望重叠率曲线

3.4 基于孪生网络的视频目标跟踪和分割统一框架

为了实现在线跟踪和更快的速度，我们采用了 Bertinetto 等人提出的全卷积孪生网络框架作为基础框架。此外，为了说明我们的方法不局限于特定的全卷

积方法，我们采用 SiamFC 和 SiamRPN 作为两个代表性算法作为实例。

3.4.1 全卷积孪生网络

SiamFC: Bertinetto 等提出使用离线训练的全卷积孪生网络作为跟踪系统的基本构建模块，该网络将示例图像 z 与较大的搜索图像 x 进行比较，以获得密集的响应图。z 和 x 分别是以目标对象为中心的 $w\times h$ 区域和以目标的最后估计位置为中心的较大区域。 两个输入由相同的卷积神经网络 f_θ 提取特征，由交叉相关得到响应图：

$$g_\theta(z,x)=f_\theta(z)*f_\theta(x) \tag{26}$$

在本文中，我们将响应映射 g_θ 中的每个空间元素称为候选窗口(RoW)的响应。例如，$g_\theta^n(z,x)$ 编码模版 z 和在 x 窗口中的第 n 个候选窗口之间的相似性。对于 SiamFC，目标位置是响应的最大值的目标位置所对应的搜索区域 x。相反，为了让每一通道编码更丰富的目标对象信息，我们将简单的交叉相关替换为分层相关操作同时产生多通道响应图。SiamFC 采用逻辑损失离线训练数百万个视频帧，我们把这个损失函数记为 L_{sim}。

SiamRPN：Li 等人利用区域建议网络(RPN)显著提高了 SiamFC 的性能，该区域建议网络允许使用可变长宽比的矩形框来估计目标位置。特别地，在 SiamRPN 中，每一通道编码一组 k 个铆点预设和相应的目标/背景分数。因此，SiamRPN 输出与分类分数并行的预测框。利用 soft L1 和交叉熵损失对两个输出支路进行训练。在下面，我们将它们分别称为 L_{box} 和 L_{score}。

SiamMask：不同于现有依赖于低保真对象表示的跟踪方法，我们认为生成每帧二值分割掩码具有重要性。除了相似度评分和矩形边界框坐标外，全卷积孪生网络的通道还可以对生成像素级二值掩码所需的信息进行编码。这可以通过使用额外的分支和损失来扩展现有的孪生网络跟踪器来实现。

我们为每一个 RoW 预测一个 $w\times h$ 二值分割，使用一个简单的两层神经网络 h_ϕ 用来学习参数，可学的参数记作 ϕ。设 m_n 表示第 n 个 RoW 的预测掩码，

$$m_n=h_\phi(g_\theta^n(z,x)) \tag{27}$$

从上式我们可以看到分割的预测是由搜索区域 x 和跟踪目标 z 共同决定的函数。通过这种方式，可以使用 z 作为参考指导分割过程，这使得算法可以跟踪任意类的目标对象。给定不同的参考图像 z，网络将为 x 生成不同的分割掩码。

损失函数。在训练过程中，每一个 RoW 都被标记为基于实值的二值标记 $y_n\in\{\pm1\}$，并与大小为 $w\times h$ 的基于实值的掩码 c_n 相关联。设 $c_n^{ij}\in\{\pm1\}$ 为像素 (i,j) 在第 n 个候选窗中的对象掩码。掩模预测任务的损失函数 L_{mask} 为所有位置

的二值逻辑回归损失:

$$L_{mask}(\theta,\phi)=\sum_n(\frac{1+y_n}{2wh}\sum_{ij}\log(1+e^{-c_n^{ij}m_n^{ij}})) \tag{28}$$

因此, 分类器 h_ϕ 由 $w\times h$ 个分类器组成, 每个指示给定的像素是否属于候选窗口中的对象。注意， L_{mask} 只考虑 RoW 中的正样本(即 $y_n=1$)。

掩码表示。与语义分割方法中的代表作 FCN 和 Mask R-CNN 不同，它们在整个网络中显式地保持空间信息，而我们用一个向量来表述目标的掩码。特别是, 在我们的算法中，这种表述对应于 17×17 个 RoW 产生的 $f_\theta(z)$ 和 $f_\theta(x)$ 逐通道的互相关。网络 h_ϕ 由两个 1×1 卷积层, 分别包含 256 个和 63^2 个通道。这使得每个像素分类器利用信息包含在整个通道中, 从而有一个对应候选窗口 x 的完整的视图, 这是消除实例之间看起来像目标这种歧义的关键。为了生成更精确的对象掩模，我们采用了自顶向下的精细化策略，它使用由上采样层和跨层连接组成的多个细化模块来合并低分辨率和高分辨率特性。

分割精细化模块。为了生成更精确的目标掩模，我们采用了自顶向下的网络结构，它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功能。图 16 左图表示了如何使用堆叠的精细化模块生成掩码。图 16 右图给出了精细化模块。

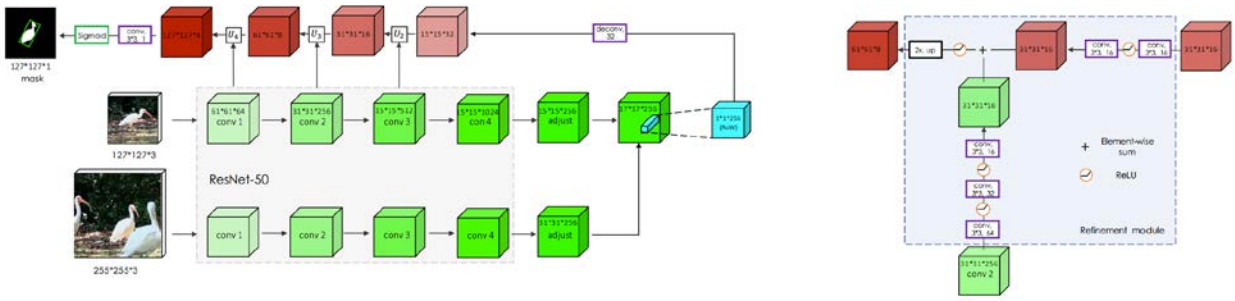


图 16. 自顶向下的分割精细化结构（左图）以及精细化模块（右图）。

两种变体。在后续的实验中，我们将 SiamFC[20]和 SiamRPN[22]的架构增加了本文提出的分割分支和分割损失 L_{mask}，得到了我们所说的 SiamMask 的两个分支和三个分支的变体。它们分别优化了多任务损失 L_{2B} 和 L_{3B}, 定义为:

$$L_{2B}=\lambda_1\cdot L_{mask}+\lambda_2\cdot L_{sim} \tag{29}$$

$$L_{3B}=\lambda_1\cdot L_{mask}+\lambda_2\cdot L_{score}+\lambda_3\cdot L_{box} \tag{30}$$

对于 L_{3B}，如果某一个 RoW 的锚点框中与目标标签框的 IoU 大于等于 0.6，则该行被认为是正样本($y_n=1$)， 否则该行被认为是负样本($y_n=-1$)。对于 L_{2B}，

我们采用与 SiamFC 相同的策略来定义正样本和负样本。我们没有搜索的上面两个公式中的超参数和简单地设置 $\lambda_1 = 32$ 和 $\lambda_2 = \lambda_3 = 1$。矩形框和得分输出的任务相关分支由两个 1×1 的卷积层组成。下图说明了 SiamMask 的两个变体。

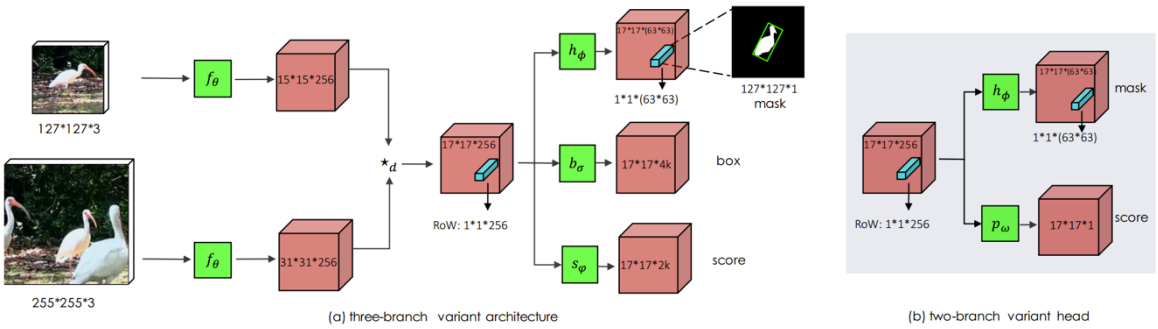


图 17. SiamMask 的示意图，左图为三分支架构，右图为两分支架构。

矩形框生成。虽然 VOS 基准测试需要二值掩码，但是典型的跟踪基准测试(如 VOT)需要一个边界框作为目标对象的最终表示。我们考虑了三种不同的策略，从一个二值掩码生成一个边界框(图 18)：(1)轴向对齐的边界矩形(Min-max)，(2)旋转的最小边界矩形(MBR)和(3)VOT-2016 中提出的用于自动边界框生成的优化策略(Opt)。我们在下节中对这三种方案进行了实验评估(表 4)。

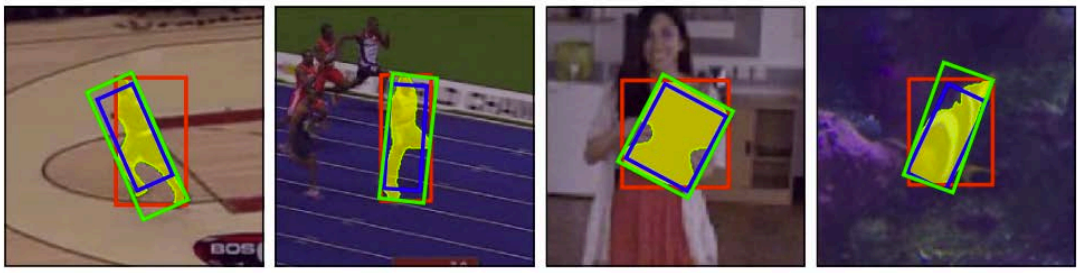


图 18. 利用二值分割结果(黄色)来生成坐标轴对齐的最小外界矩形(红色)，最小旋转外接矩形（绿色），最优比例矩形（蓝色）

3.4.2 实现细节

网络体系结构。表 3 显示了我们的骨干架构 f_θ 的详细信息。对于这两种变体，我们都使用 ResNet-50，直到第 4 阶段的最后一层卷积。为了在卷积网络深层获得更高的空间分辨率，我们使用与步幅为 1 的卷积将输出的有效步长减小到 8。此外，我们通过使用膨胀卷积来增加感受野。具体来说，我们将 conv4 的 3×3 卷积层的步幅设为 1，膨胀率设为 2。与原来的 ResNet-50 不同的是，conv4 中没有向下采样，我们还在主干上增加了一个调整层(一个 1×1 的卷积层，有 256 个输出通道)。模板和搜索区域共享从 conv1 到 conv4 的网络参数，而调整层的

参数不共享。将调整层的输出特征进行逐通道交叉相关，得到大小为 17×17×256 的相关特征图。

block	*examplar* output size	*search* output size	*backbone*
conv1	61×61	125×125	7×7, 64, stride 2
conv2_x	31×31	63×63	3×3 max pool, stride 2 $\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times3$
conv3_x	15×15	31×31	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times4$
conv4_x	15×15	31×31	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times6$
adjust	15×15	31×31	1×1, 256
xcorr	17 × 17		depth-wise

表 3. 网络主干的结构图

训练。和 SiamFC 一样，我们使用模板和搜索区域图像块分别为 127×127 和 255×255 像素。在训练过程中，我们随机抖动来对模板和搜索区域进行数据增强。具体来说，我们考虑随机平移(最多±8 个像素)和随机缩放(对于模板和搜索区域分别采用 $2^{\pm1/8}$ 和 $2^{\pm1/4}$)。对网络主干进行 ImageNet-1k 分类任务的预训练。我们使用 SGD 的第一个预阶段，在第一个 5 个周期内学习速率线性增加，从 10^{-3} 到 5×10^{-3}，然后对数递减再训练 15 个周期直到 5×10^{-4}，。我们用 COCO、ImageNet-VID 和 YouTube-VOS 来训练我们所有的模型。

测试。在跟踪过程中，SiamMask 只是每帧评估一次，没有任何在线训练。在我们的两个变体中，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。然后，在应用了每个像素的 sigmoid 之后，我们用阈值 0.5 对掩模分支的输出进行二值化。在两分支变体中，对于第一个视频帧之后的每个视频帧，我们将输出分割结果与 Min-max 框匹配，并将其作为参考来裁剪下一个帧搜索区域。在三分支变体中，我们发现利用 box 分支输出作为参考得分最高。

实验结果

在本节中，我们将评估我们的方法在两个相关任务上的效果:视觉对象跟踪(关于 VOT-2016 和 VOT-2018)和半监督视频对象分割(关于 DAVIS-2016 和 DAVIS-2017)。我们分别用 SiamMask-2b 和 SiamMask 表示我们的两个分支和三个

分支的变体。

3.4.3 视觉目标跟踪中的测试

数据集和设置。我们采用两个广泛使用的基准来评估对象跟踪任务:VOT-2016 和 VOT-2018，它们都使用旋转边框进行了标注。我们使用 VOT-2016 进行了一个实验，以了解不同类型的表述如何影响性能。在第一个实验中，我们使用了平均重叠率(IOU)和在 0.5 以及 0.7 IOU 下的平均精度（AP）。然后，我们使用官方的VOT 工具包和预期平均重叠(EAO)与 2018 年的最新算法进行比较,EAO 指标同时考虑了跟踪器的准确性和鲁棒性。

目标表示的重要性

现有的跟踪方法通常使用固定或可变的长宽比来预测坐标轴对齐的矩形框。我们分析每帧生成一个二进制掩码可以在多大程度上改进跟踪。为了提高图像的表述精度，本实验忽略了时间方面的因素，对视频帧进行随机采样。下表描述的方法是在随机裁剪的搜索补丁上测试的(随机移动在±16 像素内，缩放变形达到 $2^{\pm 0.25}$)。

	mIOU (%)	mAP@0.5 IOU	mAP@0.7 IOU
Fixed a.r. Oracle	73.43	90.15	62.52
Min-max Oracle	77.70	88.84	65.16
MBR Oracle	84.07	97.77	80.68
SiamFC [4]	50.48	56.42	9.28
SiamRPN [71]	60.02	76.20	32.47
SiamMask-*Min-max*	65.05	82.99	43.09
SiamMask-*MBR*	67.15	85.42	50.86
SiamMask-*Opt*	**71.68**	**90.77**	**60.47**

表 4. 不同矩形框表示策略在 VOT2016 上的性能对比

表 4 中,我们比较三分支变体使用 Min-max，MBR 和 Opt 方法,我们也报告 SiamFC 和 SiamRPN 代表固定和可变比例方法的结果,结合三个理想实验获得每帧真实信息和不同的表述策略作为上界。(1)固定长宽比理想实验使用每帧的真实目标区域和中心位置，但将长宽比调整为第一帧中的目标长宽比，并生成一个轴向对齐的边界框。(2) 最小外界矩形框理想实验使用旋转的理想矩形边界框的最小外接矩形来生成一个轴向对齐的边界框。(3)最后，旋转外界矩形理想实验使用旋转的最小边界矩形真值。注意(1)、(2)和(3)可以分别认为是SiamFC、SiamRPN 和 SiamMask 表示策略的性能上界。

从表 4 可以看出，无论使用何种矩形框生成策略，我们的方法都能获得最佳的 mIOU。虽然 SiamMask-Opt 提供了最高的 IOU 和 mAP，但由于其优化过程较慢，需要大量的计算资源。SiamMask-MBR 实现了 85.4%的 mAP@0.5 IOU，相对于 SiamFC 和 SiamRPN 分别提高了+29%和+9.2%。当考虑到 0.7 IOU 的精度时，我们的精度提升达到+41.6%和+18.4%，这一差距显著扩大。值得注意的是，我们的精度结果与固定长宽比的理想实验相差不远。此外，通过比较理想实验所代表的上限性能，我们可以注意到，只要简单地改变边界框的表示，就有很大的改进空间(例如，固定长宽比和旋转外界矩形之间的 mIOU 改进幅度为+10.6%)。

总的来说，这个研究显示了从对象的二值掩码中获得一个旋转的矩形策略比简单地报告轴向对齐的边界框的策略具有显著的优势。

	SiamMask-*Opt*	SiamMask	SiamMask-2B	DaSiamRPN [63]	SiamRPN [28]	SA_Siam_R [15]	CSRDCF [33]	STRCF [29]
EAO ↑	**0.387**	**0.380**	0.334	0.326	0.244	0.337	0.263	0.345
Accuracy ↑	**0.642**	**0.609**	0.575	0.569	0.490	0.566	0.466	0.523
Robustness ↓	0.295	0.276	0.304	0.337	0.460	0.258	0.318	**0.215**
Speed (fps) ↑	5	55	60	160	**200**	32.4	48.9	2.9

表 5. 在 VOT2018 上与当前最好算法的对比

在表 5 中，我们将 SiamMask 旋转外接矩形策略和 SiamMask-opt 的两个变体与 VOT-2018 上与最近发布的五个跟踪器进行了比较。除非另有说明，SiamMask 指的是我们的旋转外接矩形策略的三分支变体。SiamMask 的两种变体都实现了出色的性能和实时运行。特别地，我们的三分支变体显著优于最近的、性能最好的 DaSiamRPN，实现了 0.380 的 EAO，并以 55fps 运行。即使没有矩形框回归分支，我们的更简单的双分支变体(SiamMask-2B)也可以实现 0.334 的高 EAO，与 SA-Siam R 相当，并且优于已发表文献中的任何其他实时方法。此外，SiamMask-Opt 在 EAO 为 0.387 的情况下性能最好，但运行速度只有 5fps。这是意料之中的，因为矩形框优化策略需要更多的计算来提供更高的精度。我们的模型在精度指标表现较优，与基于相关滤波器的跟踪器 CSRDCF、STRCF 相比具有显著优势。因为 SiamMask 依赖于更丰富的对象表示，如表 5 所示。与我们类似，He 等人(SA Siam R)通过考虑多个旋转和重新标度的边界框来实现更精确的目标表示。但是，它们的表示仍然受到固定的方面比框的限制。

	VOT-2018			VOT-2016			
	EAO↑	A↑	R↓	EAO↑	A↑	R↓	Speed
SiamMask-box	0.363	0.584	0.300	0.412	0.623	0.233	**76**
SiamMask	0.380	0.609	**0.276**	0.433	0.639	**0.214**	55
SiamMask-*Opt*	**0.387**	**0.642**	0.295	**0.442**	**0.670**	0.233	5

表 6. 在 VOT2018 上与当前最好算法的对比

表 6 给出了不同矩形框生成策略下 SiamMask 在 VOT2018 和 VOT2016 的结果。SiamMask-box 是指虽然已经训练了 mask 分支，但是仍然使用 SiamMask 的 box 分支进行推理。通过使用掩码分支生成 box，我们可以清楚地看到所有评估指标的改进。

3.4.4 半监督视频目标分割评估

我们的模型一旦训练完成，也可以用于视频目标分割（VOS）的任务。在实现具有竞争力的性能的同时，不需要在测试时进行任何调整。重要的是，与典型的 VOS 方法不同，我们的方法可以实时运行，只需要一个简单的边界框初始化。

数据集和设置。我们报告了 SiamMask 在 DAVIS-2016、DAVIS-2017 和 YouTube-VOS 基准上的表现。对于这两个 DAVIS 数据集，我们使用官方的性能度量:Jaccard 指数(J)表示区域相似性，F 度量(F)表示轮廓精度。对于每个评价指标 $C \in \{J,F\}$,考虑三个统计数据:均值 C_M、召回 C_O 和衰减 C_D，这告诉我们性能随时间的增加/减少。YouTube-VOS 的最终结果 O 是四个指标的平均值: J_S 代表可见的类别重叠精度， J_U 代表不可见的类别重叠精度， F_S 代表不可见的类别边缘精度， F_U 代表不可见的类别边缘精度。

为了初始化 SiamMask，我们从第一帧提供的掩码中提取轴向包围框(最小最大策略)。与大多数 VOS 方法类似，在同一个视频中有多个对象(DAVIS-2017)。我们执行多个跟踪器进行跟踪。

在 DAVIS 和 YouTube-VOS 的结果。在半监督（semi-supervised）设置下,VOS 方法初始化一个二值分割,他们中的许多需要整合计算密集型技术用来测试。例如，训练样本增强,推理 MRF / CRF 和光学流。因此，VOS 技术通常需要几分钟来处理一个短序列。显然，这些策略使得在线适用性成为不可能。因此，在我们的比较中，我们主要集中于采用最先进的快速方法。

	FT	M	$\mathcal{J}_{\mathcal{M}}\uparrow$	$\mathcal{J}_{\mathcal{O}}\uparrow$	$\mathcal{J}_{\mathcal{D}}\downarrow$	$\mathcal{F}_{\mathcal{M}}\uparrow$	$\mathcal{F}_{\mathcal{O}}\uparrow$	$\mathcal{F}_{\mathcal{D}}\downarrow$	Speed
OnAVOS [53]	✔	✔	**86.1**	**96.1**	5.2	**84.9**	**89.7**	5.8	0.08
MSK [39]	✔	✔	79.7	93.1	8.9	75.4	87.1	9.0	0.1
MSK_b [39]	✔	✘	69.6	-	-	-	-	-	0.1
SFL [9]	✔	✔	76.1	90.6	12.1	76.0	85.5	10.4	0.1
FAVOS [8]	✘	✔	82.4	96.5	4.5	79.5	89.4	5.5	0.8
RGMP [57]	✘	✔	81.5	91.7	10.9	82.0	90.8	10.1	8
PML [7]	✘	✔	75.5	89.6	8.5	79.3	93.4	7.8	3.6
OSMN [59]	✘	✔	74.0	87.6	9.0	72.9	84.0	10.6	8.0
PLM [62]	✘	✔	70.2	86.3	11.2	62.5	73.2	14.7	6.7
VPN [22]	✘	✔	70.2	82.3	12.4	65.5	69.0	14.4	1.6
SiamMask	✘	✘	71.7	86.8	**3.0**	67.8	79.8	**2.1**	**55**

表 7. 在 DAVIS2016 上与当前最好算法对比

	FT	M	$\mathcal{J}_{\mathcal{M}}\uparrow$	$\mathcal{J}_{\mathcal{O}}\uparrow$	$\mathcal{J}_{\mathcal{D}}\downarrow$	$\mathcal{F}_{\mathcal{M}}\uparrow$	$\mathcal{F}_{\mathcal{O}}\uparrow$	$\mathcal{F}_{\mathcal{D}}\downarrow$	Speed
OnAVOS [53]	✔	✔	**61.6**	**67.4**	27.9	**69.1**	**75.4**	26.6	0.1
OSVOS [5]	✔	✔	56.6	63.8	26.1	63.9	73.8	27.0	0.1
FAVOS [8]	✘	✔	54.6	61.1	**14.1**	61.8	72.3	**18.0**	0.8
OSMN [59]	✘	✔	52.5	60.9	21.5	57.1	66.1	24.3	8.0
SiamMask	✘	✘	54.3	62.8	19.3	58.5	67.5	20.9	**55**

表 8. 在 DAVIS2017 上与当前最好算法对比

	FT	M	$\mathcal{J}_{\mathcal{S}}\uparrow$	$\mathcal{J}_{\mathcal{U}}\uparrow$	$\mathcal{F}_{\mathcal{S}}\uparrow$	$\mathcal{F}_{\mathcal{U}}\uparrow$	$\mathcal{O}\uparrow$	Speed
OnAVOS [53]	✔	✔	60.1	46.6	**62.7**	51.4	55.2	0.1
OSVOS [5]	✔	✔	59.8	**54.2**	60.5	**60.7**	**58.8**	0.1
OSMN [59]	✘	✔	60.0	40.6	60.1	44.0	51.2	8.0
SiamMask	✘	✘	**60.2**	45.1	58.2	47.7	52.8	**55**

表 9. 在 Youtube-VOS 上与当前最好算法对比

这三个表显示了如何将 SiamMask 视为在线视频目标分割的强大基线。首先，它几乎比 OnAVOS 或 SFL 等精确方法快两个数量级。其次，它与目前不采用微调的 VOS 方法具有竞争力，同时比最快的方法(即 OSMN 和 RGMP)的效率高出四倍。有趣的是，我们注意到 SiamMask 在区域相似性(J_D)和轮廓精度(F_D)方面都达到了低衰减，在 DAVIS-2016 和 DAVIS-2017 上都是如此。这表明我们的方法随着

时间的推移是稳健的，因此它适用于特别长的序列。

VOT 和 DAVIS 序列的 SiamMask 定性结果见图 19。尽管跟踪速度保持很快，SiamMask 仍然可以在有干扰的情况下生成精确的分割掩码。

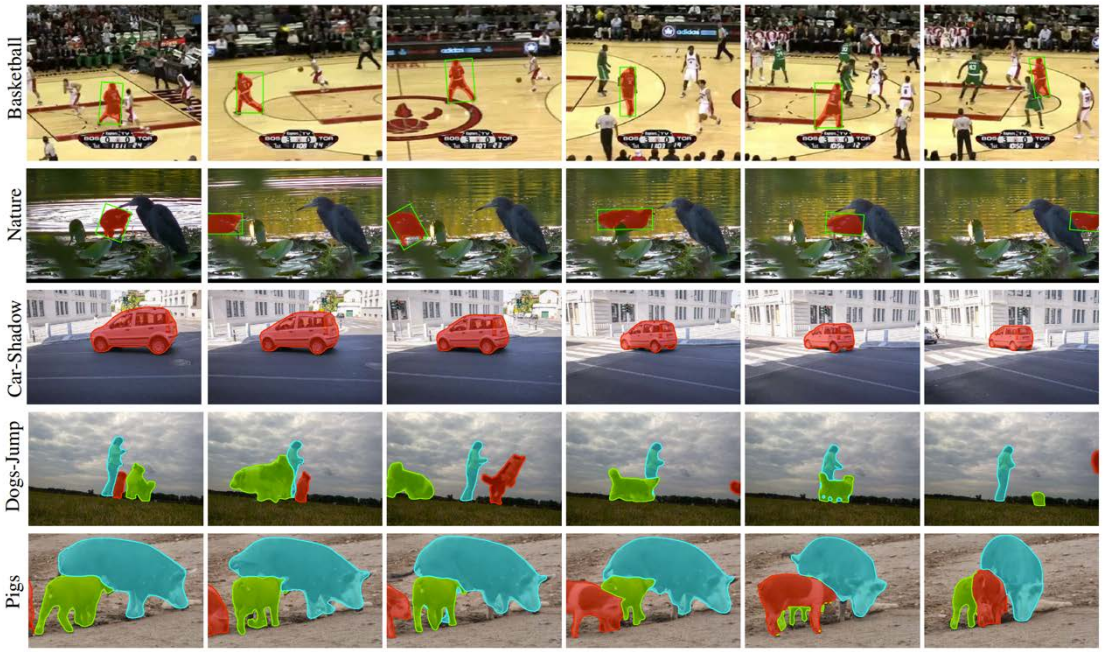


图 19. 在 VOT 和 DAVIS 数据集上的可视化结果

4、学位论文工作进度安排

从开题答辩到中期答辩期间（2018 年 10 月至 2019 年 11 月），已完成有关视频目标跟踪的注意力机制相关研究，利用编解码网络的提升网络泛化性能的研究，以及增加分割输出的多任务机制研究。

2019 年 11 月至 2020 年 1 月,继续完成孪生网络注意力机制方向的研究,完善前期工作并投稿至杂志期刊。

2020 年 1 月至 2019 年 3 月,将单目标的视频目标跟踪和分割拓展至多目标场景下，进一步提升算法性能，将论文投稿至杂志期刊。

2020 年 1 月至 2020 年 5 月,工作总结,撰写毕业论文,进行答辩。

5、已取得的阶段性成果

已发表文章：

1. **Qiang Wang**, Li Zhang, Luca Bertinetto, Weiming Hu, Philip H.S. Torr, “Fast Online Object Tracking and Segmentation: A Unifying Approach”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
2. **Qiang Wang**, Mengdan Zhang, Junliang Xing, Jin Gao, Weiming Hu, "Do not Lose the Details: Reinforced Representation Learning for Robust Visual Tracking", International Joint Conferences on Artificial Intelligence (IJCAI), 2018.
3. **Qiang Wang**, Jin Gao, Mengdan Zhang, Junliang Xing, Weiming Hu, "SPCNet: Scale Position Correlation Network for End-to-End Visual Tracking", International Conference on Pattern Recognition (ICPR), 2018.
4. **Qiang Wang**, Zhu Teng, Junliang Xing, Jin Gao, Weiming Hu, "Learning Attentions: Residual Attentional Siamese Network for High Performance Online Visual Tracking", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
5. Jin Gao, **Qiang Wang**, Junliang Xing, Haibin Ling, Weiming Hu, Stephen Maybank, “Tracking-by-Fusion via Gaussian Process Regression Extended to Transfer Learning”, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2018.

6. Zhu Teng, Junliang Xing, **Qiang Wang**, Baopeng Zhang,Jianping Fan, “Deep Spatial and Temporal Network for Robust Visual Object Tracking”, IEEE Transactions on Image Processing (TIP), 2019.
7. Zhao Yang, **Qiang Wang**, Luca Bertinetto, Weiming Hu, Song Bai, Philip H.S. Torr, “Anchor Diffusion for Unsupervised Video Object Segmentation”, The IEEE International Conference on Computer Vision (ICCV), 2019.
8. Bo Li, Wei Wu, **Qiang Wang**, Fangyi Zhang, Junliang Xing, Junjie Yan, “SiamRPN++: Evolution of Siamese Visual Tracking with Very Deep Networks”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.
9. Zheng Zhu, **Qiang Wang**, Bo Li, Wei Wu, Junjie Yan, Weiming Hu, "Distractor-Aware Siamese Networks for Visual Object Tracking", European Conference on Computer Vision (ECCV), 2018.
10. Mengdan Zhang, **Qiang Wang**, Junliang Xing, Jin Gao, Peixi Peng, Jiashi Feng, Weiming Hu, Steve Maybank, "Visual Tracking via Spatially Aligned Correlation Filters Network", European Conference on Computer Vision (ECCV), 2018.
11. Zhu Teng, Junliang Xing, **Qiang Wang**, Congyan Lang, Songhe Feng, Yi Jin, "Robust Object Tracking based on Temporal and Spatial Deep Networks." The IEEE International Conference on Computer Vision (ICCV), 2017.

参加竞赛结果：

1. Visual Object Tracking VOT2018 challenge 实时竞赛第一名。
2. Visual Object Tracking VOT2018 challenge 长时跟踪第二名。
3. DAVIS Challenge 2019 无监督组第二名。
4. 第二届 Large-scale Video Object Segmentation Challenge 视频实例分割（Video Instance Segmentation）第二名。

6、课程主要完成情况

课程学习情况					
课程名称	**分数**	**学分**	**课程名称**	**分数**	**学分**
网络数据挖掘	95	2.0	实用最优化算法及其应用	98	1.5
模式识别	84	3.0	生物特征识别	93	2.0
图像处理与分析	90	3.0	视频处理与分析	91	2.0
矩阵分析与应用	73	2.0	矩阵在信息处理中的应用	86	1.0
随机过程	98	2.0	模式识别研讨与实践	85	1.0
计算机算法设计与分析	79	3.0	中国马克思主义与当代	75	1.0
人文系列讲座	通过	1.0	中国特色社会主义理论与实践研究	75	1.0
自然辩证法概论	81	1.0	博士学位英语	68	2.0
硕士学位英语	79	3.0	统计机器学习	86	1.0
体育类公共选修课	通过	0.5	第一人称视觉	92	0.5
最优化算法理论与应用	94	2.0	模式识别与机器学习	77	2.0
学位课合计学分	29		课程学习总学分	37.5	

7、其他

（针对开题时存在的问题，本人是如何解决的，以及其他需要说明的情况）

主要参考文献

[1] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 2544-2550.

[2] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2012: 702-715.

[3] Henriques J F, Caseiro R, Martins P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.

[5] Danelljan M, Shahbaz Khan F, Felsberg M, et al. Adaptive color attributes for real-time visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1090-1097.

[6] Danelljan M, Häger G, Khan F, et al. Accurate scale estimation for robust visual tracking[C]//British Machine Vision Conference, Nottingham, September 1-5, 2014. BMVA Press, 2014.

[7] Li Y, Zhu J. A scale adaptive kernel correlation filter tracker with feature integration[C]//European conference on computer vision. Springer, Cham, 2014: 254-265.

[8] Danelljan M, Hager G, Shahbaz Khan F, et al. Learning spatially regularized correlation filters for visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4310-4318.

[9] Kiani Galoogahi H, Sim T, Lucey S. Correlation filters with limited boundaries[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4630-4638.

[10] Galoogahi H K, Fagg A, Lucey S. Learning Background-Aware Correlation Filters for Visual Tracking[C]//ICCV. 2017: 1144-1152.

[11] Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]//European Conference on Computer Vision. Springer, Cham, 2016: 472-488.

[12] Danelljan M, Hager G, Shahbaz Khan F, et al. Adaptive decontamination of the training set: A unified formulation for discriminative visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1430-1438.

[13]Danelljan M, Hager G, Shahbaz Khan F, et al. Convolutional features for correlation filter based visual tracking[C]//Proceedings of the IEEE International Conference on Computer Vision Workshops. 2015: 58-66.

[14]Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[15]Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]//Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 2017: 5000-5008.

[16]Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[C]//Advances in neural information processing systems. 2013: 809-817.

[17]Wang N, Li S, Gupta A, et al. Transferring rich feature hierarchies for robust visual tracking[J]. arXiv preprint arXiv:1501.04587, 2015.

[18]Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3119-3127.

[19]Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 4293-4302.

[20]Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[C]//European conference on computer vision. Springer, Cham, 2016: 850-865.

[21]Huang C, Lucey S, Ramanan D. Learning policies for adaptive tracking with deep feature cascades[C]//IEEE Int. Conf. on Computer Vision (ICCV). 2017: 105-114.

[22]Li B, Yan J, Wu W, et al. High Performance Visual Tracking With Siamese Region Proposal Network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8971-8980.