中国科学院大学
University of Chinese Academy of Sciences

# 硕士学位论文

## 基于自适应的深度相关滤波器的目标跟踪方法研究


作者姓名：王　攀

指导教师：韩振军　副教授

中国科学院大学电子电气与通信工程学院


学位类别：工程硕士

学科专业：计算机技术

培养单位：中国科学院大学电子电气与通信工程学院

2018 年 5 月

# Object Tracking Method Study Based on Adaptive Deep Correlation Filter

**A thesis submitted to**

**University of Chinese Academy of Sciences**

**in partial fulfillment of the requirement**

**for the degree of**

**Master of Engineering**

**in Signal and Information Processing**

**By**

**Pan Wang**

**Supervisor: Associate Professor Zhenjun Han**

**School of Electronic, Electrical and Communication Engineering**

**University of Chinese Academy of Sciences**

**May 2018**

## 中国科学院大学
## 研究生学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日　　期：

## 中国科学院大学
## 学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延迟期后适用本声明。

作者签名：　　导师签名：

日　　期：　　日　　期：

# 摘 要

视觉目标跟踪是计算机视觉的重要任务之一，其目的是给定首帧目标物体标注框后完成后续帧目标物体的状态输出。作为视频理解、车辆辅助驾驶系统、无人机目标追踪和监控领域等计算机视觉相关任务不可缺少的组件，目标跟踪具有重要研究意义。

在跟踪领域，判别的相关滤波器（DCF）是一种简单且有效的算法，它通过将训练模型转换到频域中，封闭式快速求解一个线性模型来区分图像和其变换的图像。近年来，深度卷积神经网络（CNN）在视觉应用上的成功，推动大型数据集预训练的 CNN 特征结合相关滤波器的工作在跟踪领域快速发展。本文工作是在深度学习和相关滤波器相结合基础上展开，主要贡献如下：

1、建立深度相关滤波器框架（DeepCF）。详细分析传统相关滤波器原理，由传统封闭解析式回归变换到基于深度神经网络卷积式回归，建立一种鲁棒的深度相关滤波器的外观模型，缓解了传统相关滤波器由循环或是边界效应带来的模型不稳定性问题。此外，不同于大多基于 CNN 的相关滤波器模型中将两者作为独立的部件，DeepCF 采用端到端的方式联合优化；

2、提出了自适应深度判别相关滤波器模型（adaDDCF）。由于特定类别和有限数据量训练的分类 CNN 特征，在跟踪域中对“未见过”目标表征能力不够，本文在深度相关滤波器（DeepCF）的基础上，耦合了在线特征学习策略，有效的将深度分类域特征迁移学习至适应跟踪域的特征，在线自适应的学习判别能力更强的特征，构建一个端到端的自适应深度判别相关滤波器模型（adaDDCF）。此外，本文提出并且推导一个可微分的 Fisher 判别层，其可应用于任意深度卷积网络中用于生成具有判别力的特征。

3、在 OTB2013、OTB2015 和 OTB50 三个公共数据集上进行了详细实验测试，取得优异的实验性能，对比同类框架的跟踪算法具有较好的性能。

**关键词**：相关滤波器，深度学习，特征学习，目标跟踪

# Abstract

Visual object tracking which continuously locate a target of interest in successive video frames is one of the important tasks of computer vision. As an indispensable component of computer vision related tasks such as video understanding, vehicle assistant driving system, UAV target tracking and monitoring field, the research of object tracking is of great significance.

In the field of object tracking, discriminant correlation filter (DCF) model is an efficient algorithm that learns to discriminate an image patch from the surrounding patches by solving a large ridge regression problem extremely efficiently. In recent years, the deep convolutional neural networks (CNNs) have been successfully used in the vision computer application, and that has promoted greatly the combination of the pre-trained CNN features with correlation filter. In this paper, our work is based on CNN-CF, and the main contributions as follows:

1. The deep correlation filter (DeepCF) model. On the basis of detailed analysis of the traditional correlation filter principle, we establish the deep correlation filter (DeepCF) appearance model with the convolutional regression. The DeepCF effectively alleviates the circularly shifting and boundary effect in the traditional correlation filters. In addition, different from most CNN-CF based models, CNN and CF are used as independent components, and DeepCF adopts end-to-end optimization;

2. The adaptive deep discriminative correlation filter (adaDDCF) model. The CNN trained for general image classification purpose lacks sufficient discriminative capacity for any given objects in a visual tracking scenarios. Based on the deep correlation filter, the online feature learning strategy is coupled, which can effectively transfer the pre-trianed classification CNN features to object tracking domain, and construct the adaptive deep discriminative correlation filter model (adaDDCF) with end-to-end. In adaDDCF, a convolutional Fisher Discriminative Analysis (FDA) layer, which is differentiable and thus can be implemented into any deep convolutional neural network,

updating convolutional features to scene-specific discriminative features.

3. Extensive the detailed experiments on the challenging benchmarks OTB2013, OTB2015, and OTB50 demonstrate that the proposed adaDDCF tracker outperforms many state-of-the-art trackers.

**Key Words**：Correlation filter, Feature Learning, Deep Learning, Visual Object Tracking

# 目 录

# 图目录

# 表目录

# 第 1 章 绪论

## 1.1 引言

视觉是人类与外部世界进行交互所依赖的最重要的途径。在人类与外界的交互中，有 80%以上的信息经视觉获得，而且还在不断地增加，处理这些信息需要大量的复杂运算。计算机视觉的研究目的是通过对视觉对象的表达和学习，使得计算机具备自动识别和理解视觉信息的能力，实现智能感知世界的技术。作为计算机视觉领域一个基础性研究内容，目标跟踪的任务就是连续推断视频中的目标状态，定位目标并生成轨迹，如图 1-1。计算机视觉目标跟踪技术可以帮助在海量视频中跟踪感兴趣目标，建立目标在时域上的联系，同时建立图像检测到视频分析的中间桥梁。作为计算机视觉领域的重要和关键研究内容，视觉目标跟踪在过去几十年中取得了重要的进展，并且仍然是领域研究热点问题之一，经久不息。

图 1-1 视觉目标跟踪

## 1.2 课题背景和研究意义

### 1.2.1 课题背景

近十几年来，虽然涌现出大量的视觉目标跟踪的方法，但跟踪的性能和效果却仍然面临着诸多的挑战，制约其实际应用。实际场景中面临的各种复杂因素，

如目标和场景的动态变化、背景中相似物的干扰、目标的变形遮挡、尺度变化和旋转、以及运动模糊和噪声等，导致目标跟踪依旧是计算机视觉中较为有挑战的问题。

目标的外观模型以及在帧间对目标特征进行有效地关联匹配是目标跟踪研究的主要内容。目标的特征建模是目标跟踪系统中最重要的组成部分，如果提取的特征具有较强的区分能力，那么即使使用简单的模型，仍然能达到较好的跟踪性能。传统的目标跟踪算法使用人工设计的特征来描述目标外观，例如颜色直方图特征、HOG 特征、Haar 特征、像素特征等。但是这些人工设计特征只能针对特殊的跟踪场景才能发挥出较好的性能，这就导致了基于这些特征设计的目标跟踪算法在跟踪某些特定场景中的目标时取得了较好的效果，但是跟踪另外一些场景中的目标时容易丢失目标。其原因主要是这些依靠低层手工设计特征建模的跟踪算法仅仅将目标视为一堆特征的集合而并不知道跟踪的目标是什么或属于哪一类，最终导致这些算法在跟踪复杂场景中的目标时容易失败。根据视神经科学的研究，人的视觉系统的信息处理是分层的，从低层的边到高层的目标逐层抽象。高层抽象特征包括目标的显著性、语义等丰富的信息，抽象的层次越高越容易分类。因此，在局部手工特征难以描述目标外观变化的情况下，研究基于高层抽象特征的外观建模具有较为重要的理论和应用价值。

深度学习模型因擅长提取高层抽象特征而得到了广泛的关注和研究，它的应用领域包括自然语言处理，图像处理，模式识别等。在计算机视觉领域，卷积神经网络的应用最为广泛。它的局部连接、权值共享以及池化（pooling）操作等特性可以有效地降低网络的复杂度，减少训练参数的数目，使得模型对平移、旋转、缩放具有一定程度的不变性，并具有强鲁棒性和容错能力，并且也易于训练和优化网络结构。这些优异的特性使得卷积神经网络在图像分类和识别中取得了非常优异的性能。

### 1.2.2 课题的应用领域

目标跟踪涉及到特征提取、数字图像与视频处理、模式识别以及机器学习等多个方面。同时，由于其关注对象是任意视觉目标，因此对感兴趣目标的建模，需要回答目标在哪和目标有多大两个问题，其典型应用有如下几个方面：

(1) 智能视频监控

智能视频监控中，最关心的是运动目标的行为活动。在道路十字路口，可以有效地分析行人的活动轨迹，保障公共区域安全；在小区，可以发现可疑人员行动路径，对危险可疑人物进行监控；在机场、火车站以及景点，可以通过行人目标跟踪判定特征人员运动方向。有效的目标跟踪，在高效地保证公共场所的安全的同时，可以大量节省人力物力资源，产生庞大的社会价值。

图 1-2 视觉目标跟踪常见应用场景

(2) 驾驶辅助系统

汽车已经成为现代人出行不可或缺的一个重要工具，但各种交通事故给人们带来很大威胁。在实际的车辆驾驶中，对于车辆控制决策依据主要来源于视觉，比如：交通标志、路面状况、标线和信号、障碍物等。在交通环境中，如果能有效地进行运动目标跟踪，提前预测物体的运动方向，从而可以控制车辆紧急避险并有效减小损失。谷歌、百度等互联网及人工智能公司及各大汽车厂商目前均致力于自动驾驶研究及应用，为目标跟踪技术提供了广阔的应用空间。

(3) 人机交互

目前主流的人机交互是通过鼠标和键盘的操作来输入和控制。而基于手势的识别，姿态估计，行为动作的识别的高级智能交互中，跟踪技术是不可缺少的部件。简单的单帧图像的检测识别和静态分析十分困难，很难定义行为动作类别，而通过跟踪技术可以对运动图像在时间和空间上进行更好的关联，从而得到更为准确的结果。

## 1.3 国内外研究现状

在过去几十年里，目标跟踪领域取得了相当大的进展。在这一节中，本文分

析跟踪中最重要的两个部分，目标的外观表示和框架选择。在此，本文讨论了目标跟踪相关的具有挑战性的因素。

### 1.3.1 跟踪中的目标外观表示

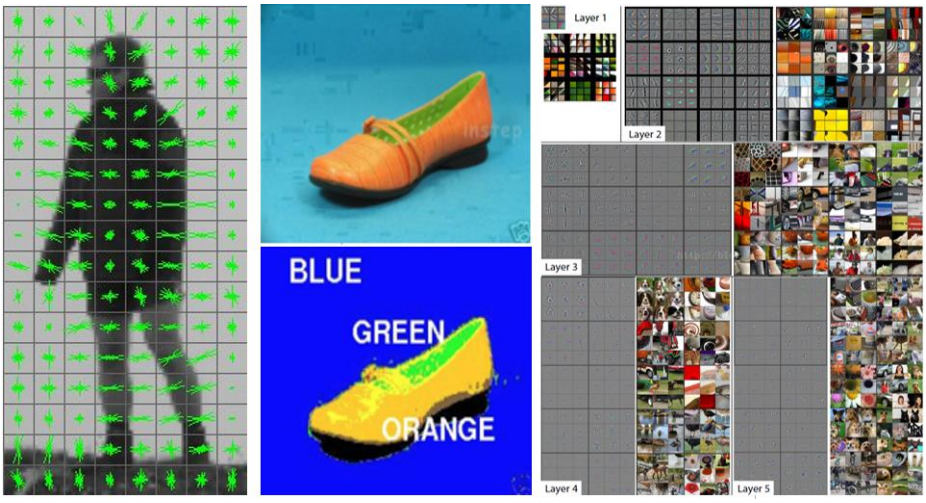



图 1-3 HOG 特征、Color Name 特征和深度特征

目标表示是视觉跟踪算法的主要组成部分之一，已有大量的目标表示方法[1]被总结。如图 1-3 为常见的几种特征表示。Lucas 和 Kanade（LK）[2]的早期研究[3,4]被广泛应用于视觉目标跟踪[5]。LK 方法不考虑目标的外观变化，因此，当目标的视觉属性发生显著变化时，其跟踪效果往往表现不佳。由 Hager 等[6]提出了一种高效的 LK 算法，并在不同光照条件下使用低维的特征表示进行跟踪。Ross 等[7]通过增量地学习低维子空间的表示来解释目标跟踪的目标外观变化。基于稀疏表示的跟踪方法[8, 9]使用了由目标和小模板组成的整体模板字典，通过求解$l_1$最小化来确定目标位置。为了处理跟踪中出现的遮挡，且提高运行时性能，后续的一些方法[10, 11, 12, 13]中引入了局部稀疏表示和协作表示方法。

基于颜色直方图的全局特征在跟踪上的应用[14, 15]较大提高跟踪性能。Comaniciu 等[14]在颜色直方图的基础上，将均值偏移算法应用于目标跟踪。Collins[16]扩展了均值漂移跟踪算法来处理跟踪目标的尺度变化。Perez 等[15]人在粒子滤波[17]中嵌入了颜色直方图，用于物体的跟踪。Birchfield 和 Rangarajan[18]不依赖于像素的统计数据，提出了一种空间图，以捕捉像素的统计特性和它们的

空间关系。局部敏感直方图[19]是通过考虑每个像素局部区域的贡献，来更好地描述跟踪物体的外观。为了利用局部的方向边缘信息，将梯度直方图（HOG）[20, 21]用于跟踪。为了融合不同类型的特征，引入了基于协方差区域描述符[22]的表示来进行目标跟踪。此外，一些方法[23, 24, 25, 37]利用局部二进制模式[27]和 Haar-like 特征[28]对目标的外观进行了建模。

依赖于简单像素统计和底层视觉空间信息的特征描述只能将跟踪性能提升到一定程度，在一些难度较大的场景，例如：遮挡，外观剧烈变化和伪目标等场景下，这种手工设计的特征往往很难长时间鲁棒的跟踪到目标。近几年来，采用学习的方式获得更强描述能力的特征成为趋势。

随着 2006 年 Hinton 在科学杂志上发表的神经网络的论文[29]和他的研究组在 2012 年 ImageNet 图像分类竞赛中取得巨大成功[30]，深度学习特征的应用进入快速增长状态。深度学习特征具有强大的特征描述能力，但是其提取计算代价非常大，很难像传统特征一样做多个尺度，而且往往也是针对一整幅图像提取特征，也很难像手工设计特征那样通过积分图快速提取某个区域的表达。除此之外，目标跟踪对实时性要求高，而深度神经网络训练需要大量样本作为输入，导致较长一段时间内，深度学习方法在跟踪领域中没有被有效利用。

### 1.3.2 框架选择

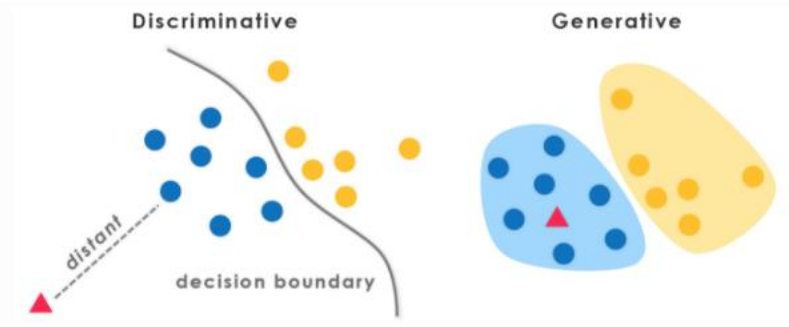



图 1-4 判别模型与生成模型[86]

基于跟踪的模型可以大体分成两类，生成模型和判别模型（图 1-4）。在生成模型[31, 32]中，跟踪是为了搜索与目标对象最相似的区域。先前，已有大量基于生成模型的方法用于搜索相似目标来估计目标的状态，例如，基于广义的 Hough 变换的跟踪算法[33]，基于稀疏的局部外观生成模型[34]，和基于梯度相似度度量算法

[35]。然而，生成模型中只考虑正样本而忽略了背景中的负样本，从而导致了在跟踪过程中由于跟踪目标外观模型的变化，经常导致跟踪失败。

在判别模型[36,37]中，同时考虑了跟踪过程中的前景和背景信息，将跟踪当作一个分类问题区别前景和背景，例如，早期的一些工作[36,37]以在线的方式训练了一个二分类器，区分前景与背景。在随后的推广中，大量的分类思想已被用于目标跟踪，如支持向量机（SVM）[36]、结构化的 SVM[38]、基于排序的 SVM[39]、提升方法（Boosting）[40]、在线多实例 Boosting [41]。为了适应外观变化，Avidan[42]在光流框架中集成了一个经过训练的 SVM 分类器进行跟踪。Collins 等[43]通过在线学习最具判别性的特征组合，在每个帧学习一个置信图，用于将目标前景与背景分离。Avidan 等[44]通过一个在线学习的弱分类器被用于确定像素是否属于目标前景或背景。在接下来的研究中，多实例学习(MIL)也被应用于跟踪[45]，算法将所有标签模糊的正和负样本都被放入包中，从而学习一种判别模型。但是在上述判别模型方法中，往往采用分类器预测跟踪目标标签，但是此约束式不能与目标跟踪实际需求的位置估计和大小估计相耦合。通过缓解这些简单分类中采样二义性的问题，基于相关滤波器模型[26]，判别相关滤波器（DCF）框架[48,49,51,54,59,60,61,64,66,67,73]被提出，其对样本采用软标签（从 0-1 之间的连续标签），而不是简单分类器学习的二元标签，在跟踪过程中通过将训练样本回归到高斯函数，从而获取样本的软标签。

但是，判别模型中，对跟踪目标的特征表示和模型仍没有很好的耦合在一起，设计的特征是完全独立于模型本身，导致很难准确地评价跟踪结果的可信度。因此，一个成功的模型应该利用生成模型和判别模型的优点[12, 27]来解释外观的变化，并有效地将前景目标前景与背景分离开来。

## 1.4 研究内容

课题的主要研究内容是基于深度的相关滤波框架的目标跟踪，如图 1-5，工作内容包括以下几个方面：

（1）深度相关滤波外观建模

本文将传统的判别相关滤波器（DCF）理解为深度网络中的卷积滤波器，通

过卷积相关滤波器建立外观模型。本文试图将传统的相关滤波器思想升级到基于深度神经网络中的深度相关滤波器。通过多层卷积滤波器替代传统浅层相关滤波器回归模型。并且，采用小尺寸的滤波器编码代替传统整体编码的相关滤波器，如 KCF[67]，HCF[60]等，其对局部特征编码可以减少背景信息的引入，从而有利于减少背景信息编码带来的干扰。同时，脱离循环矩阵特性的应用可以缓解由循环特性带来的边界效应，使得基于深度更新学习的相关滤波算法具有更好的鲁棒性。

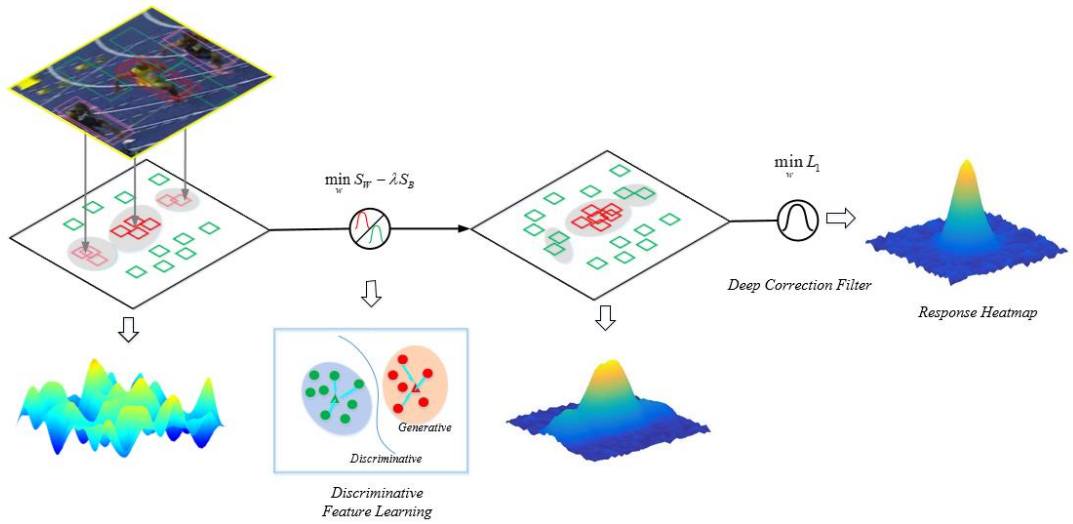



图 1-5 深度的相关滤波框架的目标跟踪框架

（2）耦合的特征学习策略

在深度相关滤波器中，进一步引入特征学习策略，其主要用于将分类任务预训练的深度分类域特征转换到适应跟踪问题的跟踪域特征。先前基于 CNN 特征的相关滤波器框架，只是简单的将相关滤波器用于预训练的 CNN 特征之上，没有深刻理解这两部分。在本文中，采用一种全新的方式，通过引入和学习一个特殊的 Fisher 判别层来学习具有判别力的特征。本文将搜索域内的前景目标和背景分成两类，Fisher 判别层通过利用在融合多层 CNN 特征上的 Fisher 判别准则对模型进行在线训练，使得前景类和背景类同类的类内散度更小，不同类别的类间散度大，以此通过这种可学习的判别特征层使前景和背景具有更好的区分性。

## 1.5 本文组织结构

第一章，绪论。论述了视觉目标跟踪的研究背景与研究意义，分析了当前视觉目标跟踪中存在的难点和常见问题，明确了本文的主要研究目的和研究内容。

第二章，相关工作。介绍了相关滤波器、深度学习、深度相关滤波器以及特征学习在目标跟踪领域的应用。首先，介绍了相关滤波器基本原理和相关研究的发展史。其次，叙述了近年来深度学习在目标跟踪领域的应用及相关跟踪算法。随后，介绍了基于深度学习和相关滤波器结合方法在目标跟踪领域的算法研究。最后，阐述了跟踪领域特征学习的思路，以及基于特征学习相关技术。通过对以上相关跟踪算法的介绍，为后续章节中关于自适应的判别的深度相关滤波器的研究进行了铺垫。

第三章，深度相关滤波器外观建模。提出了深度相关滤波器的框架。深度相关滤波器将传统判别的相关滤波器思想引入深度卷积神经网络中，代替传统封闭式求解方式回归，采用深度卷积滤波器的方式回归最终的响应。卷积滤波器采用的是小滤波器模板，和卷积扫窗操作等优点，缓解传统采用整体模板引入过多背景信息和边界效应等问题。

第四章，自适应的深度相关滤波器。在深度相关滤波器的基础上建立特征学习机制，提出了自适应的深度相关滤波器框架。为了将分类域的特征迁移学习致适合跟踪域的特征，在深度相关滤波器作为外观建模基础上耦合特征学习策略，自适应的将分类特征迁移至跟踪域特征，使得跟踪特征更具判别力。深度相关滤波器建模外观在保证跟踪速度的前提下准确定位目标。耦合的特征学习策略能够学习具有判别力的特征，能够缓解模型漂移。

第五章，实验分析。本章对深度相关滤波器外观模型和自适应的深度相关滤波器进行详细的实验验证。实验分析了深度相关滤波器和自适应的深度相关滤波器模型的有效性。并且在 OTB2015[71]、OTB2013[70]和 OTB50[71]数据集上对自适应的深度相关滤波器与当前最优秀的跟踪算法进行对比实验。为更进一步说明本文框架的鲁棒性和有效性，进一步进行了时间鲁棒性实验、空间鲁棒性实验以及 11 种属性对比实验。

第六章，总结了本文的主要内容，并对未来工作进行了展望，包括采用深度

学习进一步提高底层特征的提取能力、改进端到端的目标跟踪。

# 第 2 章 相关工作

上一章节详细阐述了目标跟踪的研究背景、意义和研究现状。本章将简单介绍与本文相关的工作，分别是，传统相关滤波器在跟踪上的应用、深度学习以及深度学习结合相关滤波在跟踪上的应用和基于特征学习的跟踪算法。

## 2.1 基于传统特征的相关滤波器

近年来，基于相关滤波(Correlation Filter)的跟踪方法[26, 48, 49, 51, 54, 59, 60, 61, 64, 66, 67, 73]由于其优异的性能，吸引了众多研究者的目光。相关滤波器通过将输入特征回归为目标高斯分布来训练滤波器模型，并在后续跟踪中寻找预测分布中的响应峰值来定位目标的位置。相关滤波器在运算中巧妙应用快速傅立叶变换和循环矩阵技巧获得了大幅度速度提升。目前基于相关滤波的研究众多，从最初的基于最小化输出误差平方和 MOSSE[75]的相关滤波器算法，到引入循环矩阵和核概念的核化相关滤波器，如 DCF[67]、CSK[26]、KCF[67]等，到后来为解决尺度变化的尺度估计的相关滤波器 DSST[64]、fDDST[66]等。本节所讨论的相关滤波器模型都是基于传统手工特征。

### 2.1.1 简单二分类的相关滤波器

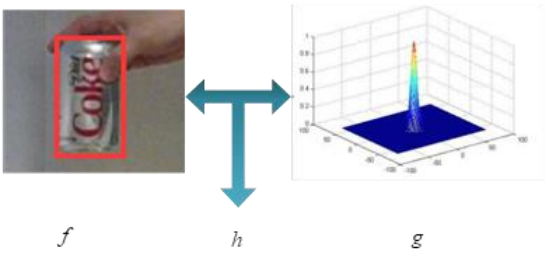



图 2-1 相关滤波器滤波器跟踪结果及卷积输出响应

2010 年，David 等[75]将相关滤波器引入视觉目标跟踪，其从信号自相关和互相关角度描述相关滤波器。相关性从直观的角度可以理解为两个信号在某个时刻的相似程度，如图 2-1。如下 $f$ 和 $g$ 两个信号的相关性表示为，

$$(f*g)(\tau)=\int_{\infty}^{-\infty} f^{*}(t)g(t+\tau)dt \tag{2-1}$$

$$(f\otimes g)(n)=\sum_{-\infty}^{\infty} f^{*}[m]g(m+n) \tag{2-2}$$

其中的 $f^{*}(t)$ 表示 $f(t)$ 的共轭。将这种相关滤波器信号相关性思想应用于视觉目标跟踪，其简单的想法为两个信号越相似，其相关值越高。在跟踪中，本文需要学习的滤波器和被跟踪物体的响应最大，即作者提出的误差最小平方和滤波器，如下，

$$g=f\otimes h \tag{2-3}$$

这里的 $g$ 表示响应输出， $f$ 表示输入的图像信号， $h$ 表示学习的滤波器。上式中如果直接求解滤波器 $h$ 复杂度极高，因此作者对以上式子进行了快速傅里叶变换，将空域上的卷积计算转化到频域的点乘操作计算，

$$G=\mathbf{F}(g)=\mathbf{F}(f\otimes h)=\mathbf{F}(f)\odot\mathbf{F}^{*}(g) \tag{2-4}$$

其中的 $\mathbf{F}(\bullet)$ 为快速傅里叶操作， $\odot$ 为点乘操作。上式可以简化为，

$$G=F\odot H^{*} \tag{2-5}$$

基于（2-5），跟踪任务中的滤波器可以求解：

$$H^{*}=\frac{G}{F} \tag{2-6}$$

在实际跟踪过程中，由于需要考虑众多的外观变化因素，如光照、形变、背景嘈杂等，为增强模型的鲁棒性，作者提出通过最小化输出误差平方和建立自适应的相关滤波器外观模型，

$$min_{H^{*}}\sum_{i}^{m}\left|F_i\odot H^{*}-G_i\right|^{2} \tag{2-7}$$

通过最小二乘思想可以求得，

$$H=\frac{\sum_i F_i\odot G_i^{*}}{\sum_i F_i\odot F_i^{*}} \tag{2-8}$$

$H$ 即为求得的滤波器模型。在作者原文中 $f_i$ 为图像采样， $g_i$ 为相对应的图像标

号，由高斯函数产生。

通过最小化输出误差平方和建立自适应的相关滤波器外观模型，将空域上的卷积相关运算转换到频域上的点乘计算使得算法变得更加简单高效，算法在普通的 PC 上运行可达到 669 帧/秒。但是 MOSSE 中只能使用单通道的灰度图作为输入，特征表达能力有限，并且模型中没有尺度的更新，对于跟踪过程中尺度变化难以应对。

### 2.1.2 核函数的相关滤波器

在基于相关滤波器框架的跟踪算法中，相关滤波器的判别能力往往会决定最终的跟踪性能。不同于 MOSSE 中通过采样稀疏的样本训练相关滤波模板得到一个线性二分类器， Henriques 等[26]提出核化循环结构的相关滤波器 CSK，密集的采样图像中相邻子窗口的循环结构用于学习一个核化正则的最小二乘的分类器，同时引入核函数将分类器投影到高维空间，从而获取一个非线性的分类器。进一步，KCF[67]作为 CSK 算法的延续，从核化的岭回归角度诠释基于循环矩阵采样的相关滤波器。CSK 和 KCF 中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 进行快速的分类器训练。引入循环矩阵实现了密集采样（图 2-2），通过密集采样将整个搜索域的特征利用起来，并不只是利用每个候选框的局部特征，从而实现了整张图片特征的提取。

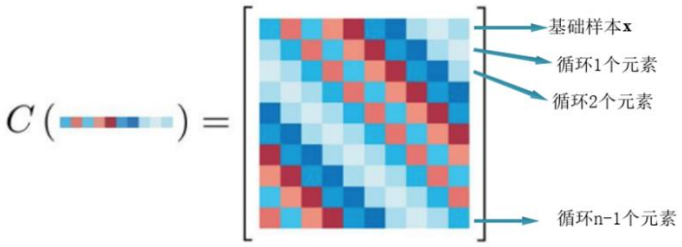



图 2-2 一维向量循环位移得到一个循环矩阵[67]

不同于 CSK，在 KCF 中，作者从岭回归的角度，通过最小二乘求出封闭解，进一步通过循环矩阵和快速傅里叶变换优化求解。训练的目标是找到模型函数 $f(z)=\mathbf{w}^T z$，最小化样本 $x_i$ 与对应回归目标 $y_i$ 的平方误差和，公式如下所示：

$$\min_{\mathbf{w}} \sum_i (f(x_i)-y_i)^2+\lambda\|\mathbf{w}\|^2 \tag{2-9}$$

其中 $\lambda$ 的为控制过拟合的正则参数，最小化方程式有闭合的解析解：

$$\mathbf{w}=(X^{H}X+\lambda I)^{-1}X^{H}\mathbf{y} \tag{2-10}$$

式中矩阵 $X^{H}$ 表示为 $X$ 的共轭转置矩阵，矩阵 $X$ 的一列表示样本 $x_i$，对应 $\mathbf{y}$ 中的回归值 $y_i$。进一步，作者通过循环矩阵思想隐式构造密集采样训练样本，假设样本为一维向量 $\mathbf{x}$，通过循环矩阵构造的训练集矩阵 $X$，

$$X=C(\mathbf{x})=\begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \\ x_n & x_1 & x_2 & \cdots & x_{n-1} \\ x_2 & x_n & x_1 & \cdots & x_{n-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_2 & x_3 & x_4 & \cdots & x_1 \end{bmatrix} \tag{2-11}$$

循环矩阵在数学上有着优越的特性，即可以被对角化：

$$X=F\,diag(\hat{\mathbf{x}})F^{H} \tag{2-12}$$

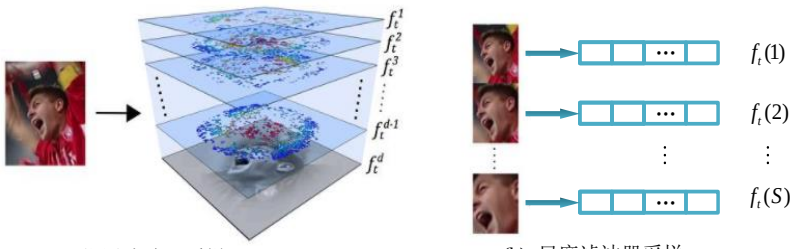
其中，$F$ 表示快速傅里叶变换矩阵，不依赖于 $\mathbf{x}$ 的常数矩阵。$\hat{\mathbf{x}}$ 表示为 $\mathbf{x}$ 的快速傅里叶变换，即 $\hat{\mathbf{x}}=\mathbf{F}(\mathbf{x})$。因此，以上模型解析式可以通过对角化特性优化表示为：

$$\hat{\mathbf{w}}=diag\left(\frac{\hat{\mathbf{x}}^{*}}{\hat{\mathbf{x}}^{*}\odot\hat{\mathbf{x}}+\lambda}\right)\hat{\mathbf{y}} \tag{2-13}$$

以一维向量为例，优化前的模型计算复杂度为 $O(n^3)$，通过循环矩阵可被对角化优化后的计算复杂度为 $O(n\log n)$，大大提升了模型求解的计算效率。

### 2.1.3 可变尺度的相关滤波器

为了解决 MOSSE、KCF、CSK 等算法中尺度不敏感问题，在 DSST[64]、fDDST[66]中使用了位置相关滤波器和尺度相关滤波器，位置滤波器解决跟踪过程中的目标位置的确定，尺度相关滤波器评估目标运动过程中的尺度变化。尺度相关滤波器作为独立于目标位置确定组件，用于目标尺度的估计，如图 2-3。




(a) 位置滤波器采样　　(b) 尺度滤波器采样

图 2-3　DSST 中的位置滤波器与尺度滤波器采样[66]

在训练尺度滤波器时，通过对目标区域不同尺度的缩放采样构成尺度金字塔，多尺度金字塔区域训练得到相关滤波器模型用于预测目标尺寸。作者在论文中采用了 33 个尺度，利用 1 维的尺度滤波器进行目标样本的尺度选择：

$$a^n w \times a^n h, n \in \left\{ -\frac{S-1}{2}, \ldots, -\frac{S-1}{2} \right\} \tag{2-14}$$

其中，$a = 1.02$ 为尺度缩放因子，$w$ 和 $h$ 分别为上一帧中样本的宽与高，$S = 33$ 表示尺度变换空间大小。

fDSST 在 DSST 基础上进一步减少计算量，选取 17 个尺度采用相关性插值扩展为 33 个尺度。同时，采用 PCA 降维将高维特征降维到低维，从而极大提高了跟踪效率。虽然基于尺度可变的相关滤波器在尺度上实现一定的自适应性，但是对于外形快速变化引起的跟踪目标尺度快速变化仍然不敏感，且循环矩阵产生的边界效应无法消除。

### 2.1.4　传统特征相关滤波器框架小结

基于相关滤波器框架的算法将跟踪视作为一个分类问题，通过相关滤波器模型从背景中区分目标。对于初始给定帧，依据给的兴趣区域训练相关滤波器模型，对于后续帧目标位置的预测流程是：对搜索域进行特征描述，对特征进行 Cosine 窗函数滤窗，在空域上抑制背景信息，随后做快速傅里叶变换将结果与相关滤波器模型做点乘操作，得出的结果经过快速傅里叶逆变换后即为目标响应热图，取响应热图最大响应位置即为目标的预测位置。

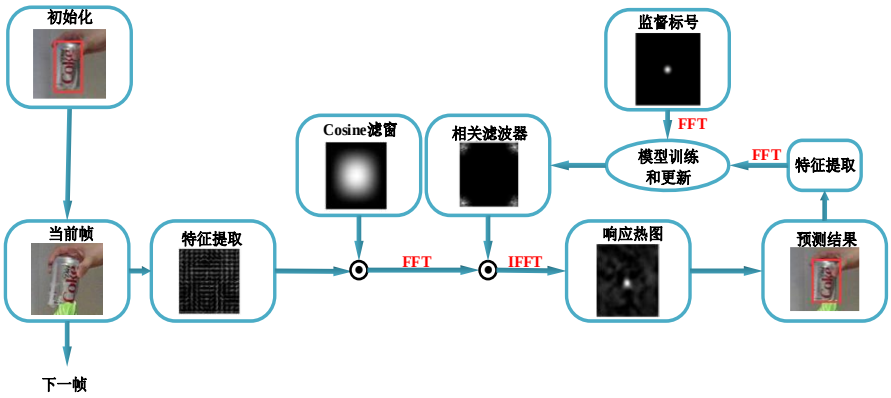



图 2-4 相关滤波器框架流程图

| 算法 | FPS | 特征 | 尺度可变 |
|---|---|---|---|
| MOSSE | 669 | 单通道灰度特征 | 否 |
| CSK | 362 | 单通道灰度特征 | 否 |
| KCF | 172 | 多通道 HOG 特征 | 否 |
| DCF | 292 | 多通道 HOG 特征 | 否 |
| CN | 152 | 多通道 ColorName | 否 |
| DSST | 65 | 多通道 HOG 特征 | 是 |
| fDSST | 86 | 多通道 HOG 特征 | 是 |

表 2-1 几种传统相关滤波器算法 FPS、特征和尺度变化敏感统计表

在整个基于传统手工特征的相关滤波器的框架中，最关注的两个方面分别是特征表达和模型优化。最初的 MOSSE 算法在模型方面简单的训练二分类器，且只适用于单通道的灰度特征，目标的外观表达能力较差且模型鲁棒性弱。CSK 在 MOSSE 基础上引入循环矩阵密集采样训练样本进一步提高模型的鲁棒性，此外引入核技巧将线性空间特征映射到高维空间使得模型的可分性更强。随后的 KCF 在 CSK 基础上将只适用于单通道灰度特征的模型扩展到多通道的 HOG 特征，同期的 CN[76]算法将单通道的特征扩展到多通道的 Color Names 颜色特征。DSST 和 fDDST 为解决以上相关滤波框架中尺度不敏感问题

提出尺度可变的相关滤波器。常见的传统相关滤波器情况统计表见表 2-1。

## 2.2 基于深度学习的跟踪算法

在深度学习被广泛使用之前，研究人员就开始从机器学习的角度入手开展针对目标跟踪的研究。这些研究通过从图像中提取手工特征，然后使用机器学习算法对目标跟踪进行建模。然而，手工特征往往针对某些场景的跟踪效果非常好，但是对于复杂场景跟踪的结果较差，缺乏泛化能力。而且随着跟踪测试集内的视频数量越来越多，种类越来越丰富，基于手工特征结合分类器建模的传统目标跟踪算法已经越来越难以适用于复杂的跟踪场景。卷积神经网络是深度学习模型的一种，它能够提取目标的抽象语义特征，这些特征的泛化能力强，在跟踪未指定类别的物体时能够比手工特征更好地描述目标，因此越来越受到研究者的重视。

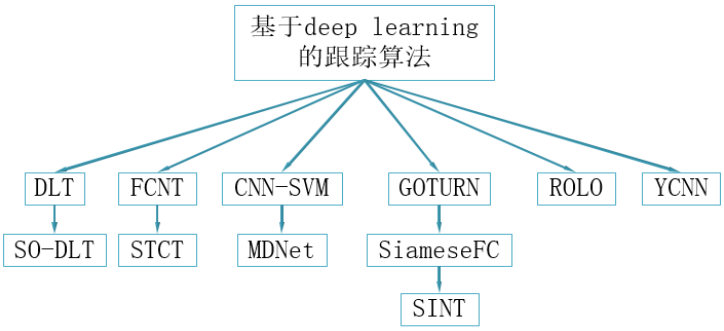



图 2-5 几种深度学习的跟踪分类

将卷积神经网络模型应用到目标跟踪领域首先要解决两个问题。第一个问题是模型的训练。训练深度模型需要大量的样本，而目标跟踪中通常只给出第一帧的目标。为了解决这个问题，深度学习跟踪算法一般会借助于图像识别或图像分类领域中的数据集来预训练模型参数，如 ImageNet 数据集等。第二个问题是模型的更新。在跟踪过程中，目标的外观会不停地改变，如何根据新跟踪到的目标在线更新已训练出的模型是一个重要的问题。没有自适应更新模块的目标跟踪算法属于静态跟踪算法，这种算法在跟踪外观变化较大的目标时容易产生漂移。另外，自适应更新算法的速度决定着跟踪的速度，只有在跟踪过程中快速更新模型才能够开发出实时的目标跟踪算法。

2010 年，Fan 等[77]首次将卷积神经网络应用到人脸跟踪上，基于采集到的人

脸数据库训练网络，该算法能较好的用于跟踪人脸。在此基础上，2013 年，Wang 等[78]使用 80 万张图片离线训练了深度神经网络，首次将深度神经网络应用到任意目标的跟踪上（而不仅仅是只跟踪人脸）。以上两种算法仅仅把卷积神经网络和深度神经网络作为特征提取器使用，属于静态跟踪算法，是研究者将深度学习模型应用于目标跟踪领域的尝试。

2015 年，SO-DLT[79]算法使用额外的数据集离线预训练卷积神经网络模型。在初始化跟踪目标时，先判断目标所属分类，再根据目标类别通过微调参数使得模型更适合检测出场景中的该类目标。同年，FCNT[58]算法深入研究了从 ImageNet 集中训练的卷积神经网络模型，并指出模型中不同水平的卷积层刻画了目标不同方面的特征。其中高层网络编码了更多的细节信息，这些信息能够将目标与背景分开；将不同层次之间的特征组合在一起能够更好地描述目标外观。CNN-SVM[63]算法使用额外图像数据集预训练卷积神经网络模型，它将网络中隐含层的输出作为特征描述子，然后将这些特征描述子输入到支持向量机中训练，最后在跟踪过程中根据分类结果确定目标，并在线更新分类器。2015 年是基于卷积神经网络的目标跟踪算法平稳发展的一年，以上三种算法已经不仅仅把卷积神经网络单纯的当成一种特征提取的黑盒算法，而是更深入的分析了网络提取特征的结构和特点。另外，这些算法都加入了在线更新模型模块，在跟踪精度上较 2013 年之前的算法有了很大提高。

STCT[53]中指出由于目标跟踪的训练样本单一，使用预训练的深度模型往往会陷入过拟合。所以提出将卷积神经网络分成多个基础学习器，每一个基础学习器使用不同的损失函数进行训练，以此来减少相关性和避免过拟合。最后将多个基础学习器集成，共同决策目标的状态。MDNet[52]将预训练的卷积神经网络拆分成多个共享层。在跟踪过程中，将多个共享层组成二分类的神经网络来分类目标。TCNN[68]使用树形结构管理目标多种外观模型，其使用多个卷积神经网络表示目标的外观，在确定目标时，根据多个网络输出的结果联合预测目标的状态。SANet[80]指出大部分基于卷积神经网络的跟踪算法将跟踪问题看做一个分类问题。由于训练的卷积神经网络的更关注于类间的区分，这些算法对场景中的相似物分离开，以此来排除相似物的干扰。ROLO[69]算法将跟踪问题建模成回归问题

而不是分类问题，在此基础上提出了一种空间监督的循环卷积神经网络。Tao[81]提出了一种没有模型更新模块、没有遮挡检测、没有融合算法的跟踪算法。它首先训练了一个 Siamese 神经网络用于提取目标特征，然后根据当前帧的目标和新一帧的目标训练一个匹配函数，根据匹配函数来判断目标的位置。GOTURN[50]是一个静态目标跟踪算法，它的训练数据来自于视频和图片。它将视频中目标的外观和其运动轨迹作为训练数据，训练一个分类 CNN 网络，跟踪速度可以达到高于 100 帧每秒。SiameseFC[82]是一个静态的目标跟踪算法。文中指出为了跟踪感兴趣的视觉目标，基于神经网络的目标跟踪算法必须在跟踪过程中使用批量梯度下降法去调整网络的权值，这极大的降低了跟踪的速度。文中使用一种 Siamese 神经网络，在跟踪过程中模型不进行更新，仅仅检测场景中的目标位置，在保证跟踪精确度的情况下能够达到实时目标跟踪。

相对于只是简单使用卷积神经网络的跟踪算法，2016 年的研究更倾向于通过剖析网络的内部结构，充分利用网络层与层之间的特征进行建模。此外，一些算法已经关注在保证跟踪精度的前提下提高跟踪速度，这标志着基于卷积神经网络的目标跟踪算法逐渐走向成熟。

## 2.3 深度学习和相关滤波器相结合的跟踪算法

之前的跟踪算法[14,16,18,20,24,67]依赖手工的特征来描述特定场景的目标，只能将跟踪问题局限于一定程度。其主要的原因在于手工制作的特性无法捕获目标的语义信息，因此，不能达到鲁棒目标外观变化的描述。深度神经网络作为各种计算机视觉任务的替代解决方案而受到关注。深度神经网络表征一个多层神经网络架构，可以有效地捕获描述原始数据的复杂层次结构。特别是，卷积神经网络（CNN）具有深层次的结构，深度卷积特征是分层的表示，即一个深层编码了更多的语义特征，并且一个浅层的信息包含了局部的细节信息[60]。

本小结对相关滤波器结合深度学习的算法从简单应用深度特征、解决边界效应、减少噪声的采样、精确定位、提高跟踪效率等方面做详细阐述。

### 2.3.1 基于静态深度特征的相关滤波跟踪算法

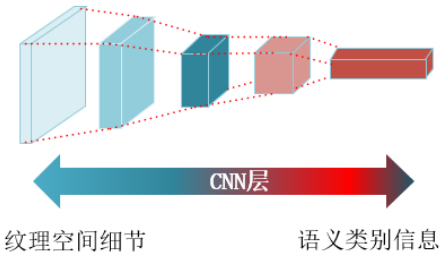



图 2-6 浅层包涵丰富的细节信息且深层包涵丰富语义信息

HCF[60] 在分析了深度预训练的分类网络中语义类别信息在更深层，且空间纹理信息在浅层的基础之上，学习了多个分层的线性相关滤波器，并使用一种较由粗到精的搜索策略定位目标。算法中使用的基本框架是 KCF 算法，将原 KCF 中的算法使用的手工 HOG 特征替换成基于 ImageNet 预训练的分类网络特征，并且作者探究了低层特征有较高的分辨率能够对目标进行精准的定位，高层特征包含更多的语义信息，能够处理较大的目标变化和防止跟踪器漂移，能够对目标进行定位。大部分的算法都只是用到了深度网络的最后一层特征，这一层的特征其实是具有一定的偏差性的，对于高层视觉识别问题，这些特征提供了有效的语义信息。但是跟踪并不是识别其语义类别，目的是定位物体的位置，仅仅用最后一层的高级语义特征，并不是最优的选择，为此在 HCF 算法中利用神经网络的各个层的特征，联合起来表示所要跟踪的物体。同时，在各个层特征自适应学习相关滤波器，而不必去进行重复采样。算法在 OTB100 上测试达到优异的跟踪性能。但是这种采用传统 KCF 的框架依旧不能解决由循环矩阵带来的边界效应，在背景嘈杂并且伪目标较多的情况下依旧会导致跟踪失败。

### 2.3.2 边界效应抑制的相关滤波器跟踪算法

2016 年，Danelljan 等[61]提出的 DeepSRDCF 为解决 KCF/DCF 中由循环特征带来的边界效应，对目标函数正则化项进行改进，使用空间正则化惩罚来改进效果。此外，文章探究了深度卷积特征对跟踪的影响，并且证明了相对于深层特征，浅层的卷积特征提供了一个更好特征表达和跟踪性能。

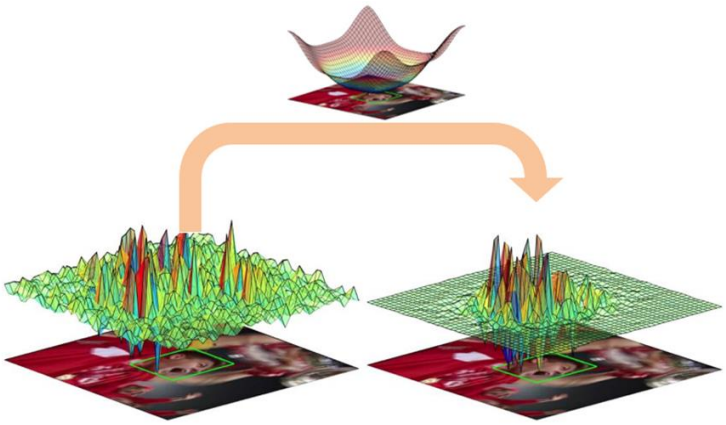


图 2-7 deepSRDCF 通过空间正则化来提高跟踪模型[61]

DeepSRDCF 算法通过一个高斯分布的空间惩罚因子，对不同位置加入不同权重的惩罚（空间正则化），最终基于空间正则化提高跟踪模型的质量。对空间正则化和输出进行可视化，在边界处的输出被明显抑制了，如图 2-7 所示。DeepSRDCF 将空间正则化的表达融合进损失函数。传统的 DCF 优化式为：

$$\varepsilon_t(f)=\sum_{k=1}^{t}\alpha_k\left\|S_f(x_k)-y_k\right\|^2+\lambda\sum_{l=1}^{d}\left\|f^l\right\|^2 \tag{2-15}$$

其中 $S_f(x)=\sum_{l=1}^{d}x^l * f^l$ 为最终的响应值， $f$ 为滤波器。SRDCF 在 DCF 约束式上引入了空间权重惩罚 $w$，公式 2-15 可以修改成：

$$\varepsilon_t(f)=\sum_{k=1}^{t}\alpha_k\left\|S_f(x_k)-y_k\right\|^2+\sum_{l=1}^{d}\left\|w\cdot f^l\right\|^2 \tag{2-16}$$

其中的惩罚项 $w$ 满足高斯分布，其作用是越靠近边界惩罚因子越大，如图 2-7 所示，通过归一化后的约束式变为：

$$\varepsilon_t(\hat{f})=\sum_{k=1}^{t}\alpha_k\left\|\sum_{l=1}^{d}\hat{x}_k^l\cdot\hat{f}^l-\hat{y}_k\right\|^2+\sum_{l=1}^{d}\left\|\frac{\hat{w}}{MN}*\hat{f}^l\right\|^2 \tag{2-17}$$

### 2.3.3 自适应净化训练样本的深度相关滤波器跟踪算法

为改进了样本和学习率问题，Danelljan 等[54]在 SRDCF 的基础上提出 SRDCFdecon。先前的相关滤波都是固定学习率线性加权更新模型，虽然这样比较简单不用保存以前样本，但在定位不准确、遮挡、背景扰动等情况会污染模型导致漂移。SRDCFdecon 选择保存以往样本(图像块包括正，负样本)，在优化目标函数中添加样本权重参数和正则项，采用交替凸搜索，首先固定样本权重，基

于高斯-塞德尔方法迭代优化模型参数，然后固定模型参数，基于凸二次规划方法优化样本权重。

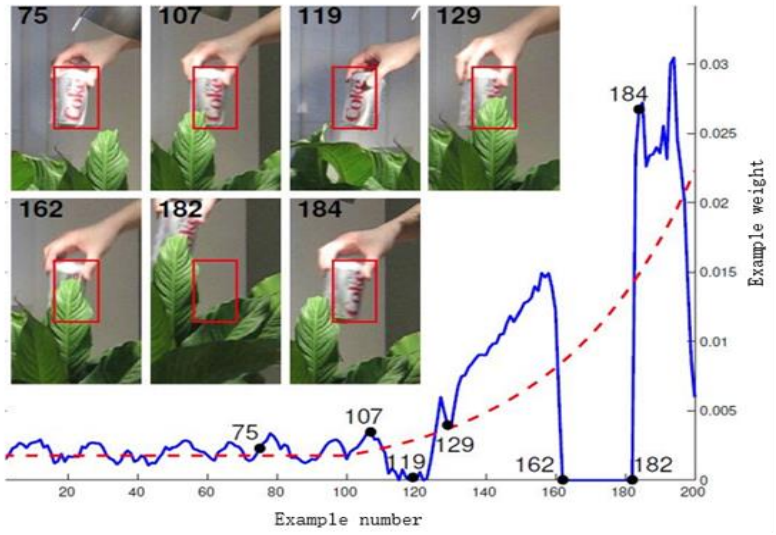



图 2-8 SRDCFdecon 自适应净化训练集合的 SRDCF 跟踪框架[54]

文中采用较为直观的思路，可信度越高的训练样本，给较高的权重，相反则降低。先前算法最简单的方法就是通过设定一个阈值判断作为采样依据，低于阈值的样本认为质量太低，直接舍弃，相反则保留。本文作者的方法是，通过训练得到样本的权重用于评判样本的质量。作者在 SRDCF 的基础上设计了以下约束方程式：

$$J(\theta,\alpha)=\sum_{k=1}^{t}\alpha_k\sum_{j=1}^{n_k}L(\theta;x_{jk},y_{jk})+\frac{1}{\mu}\sum_{k=1}^{t}\frac{\alpha_k^2}{\rho_k}+\lambda R(\theta)$$

$$\text{满足}\begin{cases}\alpha_k\geq 0,k=1,\ldots,t\\ \sum_{k=1}^{t}\alpha_k=1\end{cases} \tag{2-18}$$

其中，$\alpha_k$ 是每个样本的权重，第二项是样本权重的正则项，包括自适应度和先验样本权重。优化式求解过程为两步法，固定一个式求另一个，交替求解。利用样本的质量来训练样本的权重，提高有较高质量的样本的影响，降低较差样本的影响。将样本的质量融合到现有的损失函数约束式中，从而达到同时训练求解。和 SRDCF 相比，在 OTB100 数据集测试，OPE 性能提升约 3%（从 60.5%提升到 63.4%）。

### 2.3.4 基于连续卷积操作的相关滤波器跟踪算法

传统的判别相关滤波器（DCF）在跟踪方面取得较好的性能，其成功的关键在于能够有效地利用现有的大量负样本数据，包括利用循环矩阵转化训练样本。然而，现有的 DCF 框架仅限于单一分辨率的特征图，基于这一点 Danelljan 等[51]提出学习一种连续卷积操作（C-COT）用于相关滤波器框架，文章从传统的 DCF 框架出发，引入了连续卷积滤波器，利用隐式插值模型对连续空间域内的学习问题进行求解。此外，该方法实现的亚像素定位，这对于精确的特征点跟踪任务至关重要。

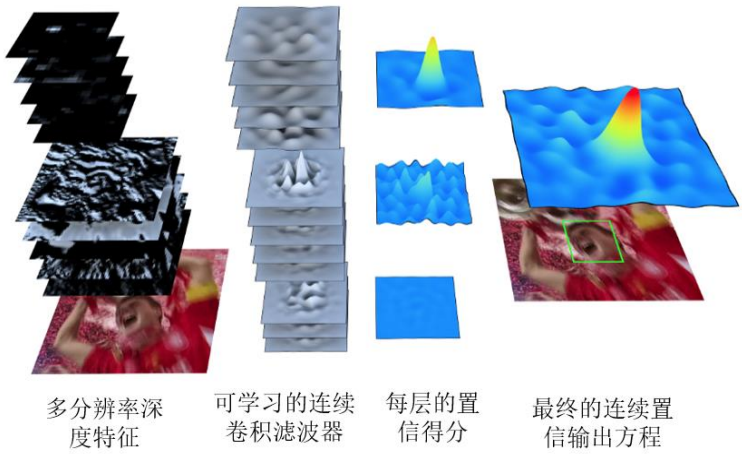



图 2-9 连续的卷积滤波操作用于多分辨率深度特征[51]

由于目标尺度变化，单个分辨率的特征图对最终输出结果存在较大扰动，将多分辨率的特征图进行融合用于更精确的定位，并且达到更鲁棒的跟踪。上图在最左边的是作者选取的特征，C-COT 基于 ImageNet 预训练的 VGG-Net 来提取特征，图中第一列即为作者使用了原始的彩色图像和两个卷积层的输出通道作为特征的示例。图中第二列为通过训练得到的连续卷积操作滤波器，每个通道对应一个滤波器，原始图像是彩色，有三个通道，对应了三个滤波器，使用这些滤波器对得到的特征图进行卷积操作就得到了第三列的响应图。将第三列的响应图进行加权平均，就得到了第四列的结果，响应图极大值位置就是预测的目标位置。

基于一维的特征作为示例，针对在连续空间域中构建学习问题，对训练样本建立了一个插值模型，对每个通道特征 $x^d$ 定义了一个插值算子 $J_d$，

$$J_d\{x^d\}(t)=\sum_{n=0}^{N_d-1}x^d[n]b_d(t-\frac{T}{N_d}n) \tag{2-19}$$

其中，插值函数$J_d\{x^d\}$可理解为插值函数$b_d$平移后叠加构成的函数，$x^d[n]$表示为特征通道$d$中第$n$维特征值，其作为相应位移函数的权值。卷积算子可以定义为：

$$S_f\{x\}=\sum_{d=1}^{D}f^d*J_d\{x^d\},x\in\chi \tag{2-20}$$

其中，每一个特征通道可以通过对应通道算子进行插值，随后与相应的卷积滤波器进行卷积。通过将计算转换至连续域得到更精确像素级别解。

假设，训练集中有$m$个训练样本$(x_j,y_j)_1^m\subset\chi\times L^2(T)$，求解卷积滤波器$f=(f^1,...,f^D)$通过以下损失函数约束，

$$E(f)=\sum_{j=1}^{m}\alpha_j\left\|S_f\{x_k\}-y_j\right\|^2+\sum_{d=1}^{D}\left\|w\cdot f^d\right\|^2 \tag{2-21}$$

C-COT 提出的方案可以有效地集成多分辨率的深度特征，从而在多个目标跟踪基准上取得优异的结果。

### 2.3.5 基于高效卷积操作的相关滤波器跟踪算法

随着基于判别相关滤波器的跟踪性能的不断提高，，跟踪模型变得越来越复杂，大量参数的引入，造成了过拟合的风险。为缓解计算复杂度和过度拟合问题，同时提高了速度和性能，高效的卷积操作（ECO）[48]做了以下三个工作：1）重新分析 DCF 的核心公式，并引入了一个因数分解的卷积运算符，极大地减少了模型中参数的数量。2）设计训练样本分布的紧凑生成模型，用于降低记忆和时间复杂度，同时提供更好的样本多样性。3）采取保守的模型更新策略，增强模型鲁棒性，降低复杂性。

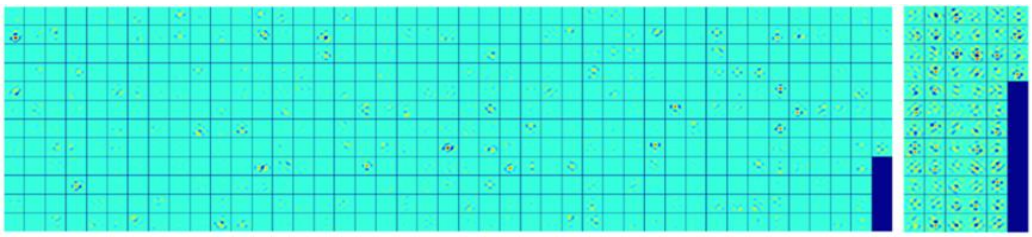



图 2-10 可视化 C-COT 和 ECO 最后一层可学习的滤波器

在 C-COT 中，作者采用了 CNN、HOG 和 CN 三种特征组合，导致每次更新模型时需要更新学习的参数多达$10^6$级，影响模型效率。在实际跟踪问题中的训练样本匮乏，且前景和背景变化较小场景下，高维度的学习参数很容易引起过拟合。文中在基于 C-COT 基础上采用因式分解卷积操作降低特征维度，新的卷积算法定义为，

$$S_{pf} = P_f * J_x = \sum_{c,d} p_{d,c} f^c * J_d\{x^d\} = f * P^T J\{x\} \tag{2-22}$$

其中，$d$ 为因式分解前特征的维度，$c$ 为降维后的特征维度，$P_f$ 是卷积滤波器组合矩阵，$J_x$ 是映射到连续空间域的卷积特征。将新的卷积算子引入先前的损失函数中，

$$E(f,P) = \left\| J\{x\}P\bar{f} - \bar{y} \right\|^2 + \left\| \sum_{c=1}^{C} \bar{w} * \bar{f} \right\|^2 + \lambda \|P\|^2 \tag{2-23}$$

作者采用用 Gauss-Newton 和 Conjugate Gradient 共轭迭代方法对公式 2-23 进行优化求解。通过在四个数据集（VOT2016，UAV123，OTB100，和 TempleColor）上测试实验，ECO 跟踪器获得了相对于原始模型 20 倍的加速。此外，使用手工特性的变种 ECO，在单 CPU 上运行达到 60FPS，在 OTB100 上平均重叠率获得 65.0%的优异性能。

## 2.4 基于特征学习的相关工作

一般来说，目标跟踪方法有三个核心任务:1)兴趣目标的数学建模；2)在新的一帧中搜索该模型的最佳位置；3)根据新获得的目标位置更新模型。所有这三个

任务的性能高度依赖于基础目标区域如何被建模成一个数学形式，即特征表示。

2005 年，Collins 等[43]一个在线的特征选择机制，通过对跟踪过程中多个特征的评价选择出一个最具判别性的特征集合来提高跟踪效果。文中，假设能将目标和背景区分开的特征是跟踪时最有效的特征。给定一个特征种子集合，通过计算目标和背景两类的条件采样密度的对数比率，形成一个新的特征来解决目标/背景的判别任务，通过提取目标和背景区域直方图，建模目标和背景的概率比率形成一个权重图。使用两个类的方差比率来对新特征的判别能力排名，即将背景和目标区分开的能力。这个特征选择机制嵌入在一个 mean-shift 跟踪系统中 使用排名最靠前的几个特征作为目标的外观描述。这种方法的限制，只能比较单通道特征，对于 HOG 这类的高维特征很难处理。

2011 年，Hare 等[38]提出了一种基于结构化输出预测的自适应视觉目标跟踪框架（Struck）。先前的基于分类的跟踪问题，将跟踪作为分类任务，利用在线学习技术更新模型。但是要实现这些更新，需要将估计的目标位置转换为一组标记的训练示例，而且无法得知怎样才能最好地执行此中间步骤。此外，分类器(标签预测)约束式并没有显式地耦合到跟踪器的约束中(目标位置的估计)。从学习的角度来看，作者在 Struck 模型中充分利用了 SVM 的分类优势，与之前基于分类的方法不同，算法并不依赖于一种启发式的中间步骤来生成标记的二进制样本来更新分类器，这通常是跟踪过程中的错误来源。文中使用了一个在线结构化的输出 SVM 学习框架，使它很容易合并图像特征和内核。为了防止支持向量数量的无限制增长，并满足实时性能，模型中引入了一个用于在线结构输出 SVMs 的预算维护机制。

2016 年，Li 等[85]通过深度神经网络卷积滤波器学习判别的特征表示（DeepTrack）。手工定义的特征表示需要专业知识，并且需要进行手动调整，因此这种方式的跟踪框架可以说是目标跟踪的限制因素之一。在 DeepTrack 中，提出了一种新颖的解决方案，在跟踪过程中自动重新学习最有用的特征表示，以便准确地适应外观变化，姿态和变化，同时防止出现拖延和失败。模型使用多个卷积神经网络（CNN）的候选库作为目标物体不同实例的数据驱动模型。其中，每个 CNN 都维护一组特定的卷积，其作用是使用所有可用的信息将目标区域与其

周围背景区分开来。这些内核在用相应的 CNN 初始化一个实例进行训练后，在每一帧以在线方式进行更新。在某一帧时，池中最置信度较高的模型被选中来评估目标物体的状态。具有最高得分的响应被指定为当前检测窗口，并且所选模型使用优化结构损失函数的来进行训练。这种基于 CNN 的在线对象跟踪器采用了 CNN 架构和结构损失函数来处理多个输入提示和类特定的跟踪。

特征描述作为目标跟踪中最基础的环节也是最重要的环节，对后续目标外观的建模至关重要。虽然有大量现有方法探讨特征学习，但自动寻找一组最佳的判别性特征表示以获得最稳健和最准确的跟踪性能仍然是一个较为困难的问题。在早期基于手工特征的跟踪方法[14,16,18,20,24,67]中，特征是手动定义和组合的，即使这些方法在单个数据集上报告令人满意的结果，手工选择特征表示也会限制视觉跟踪的性能。例如，当光照条件良好时可能具有区别性的颜色直方图特征当物体在阴影下移动时可能变得无效。近年来，深度神经网络在计算机视觉上的应用备受关注[29,83,84]。深层神经网络可以有效地捕获描述原始数据的复杂的语义信息，本文将从特征学习角度设计一种自适应的特征学习机制，并且结合深度的相关滤波器作为前景建模建立一个端到端的跟踪模型。

## 2.5 本章小结

本章节对相关滤波器、深度学习、深度相关滤波器以及特征学习在目标跟踪领域的应用进行了简单的介绍。第一节介绍了相关滤波器基本原理和相关研究的发展史。第二节叙述了近年来深度学习在目标跟踪领域的应用及相关跟踪算法介绍。第三节介绍了基于深度学习和相关滤波器结合方法在目标跟踪领域的算法研究。最后一节阐述了跟踪领域特征学习的思路，以及基于特征学习相关技术。通过对以上相关跟踪算法的介绍，为后续本文中关于自适应的判别的深度相关滤波器的研究提供引导。

# 第 3 章 深度相关滤波器外观建模

上一章分别从传统相关滤波器、深度相关滤波器、特征学习等几个方面介绍了与本文相关的技术。本章将对深度相关滤波器的外观建模进行研究，从深度特征表达分析，解析式的相关滤波，到基于深度学习的相关滤波器的外观建模进行完整的论述。

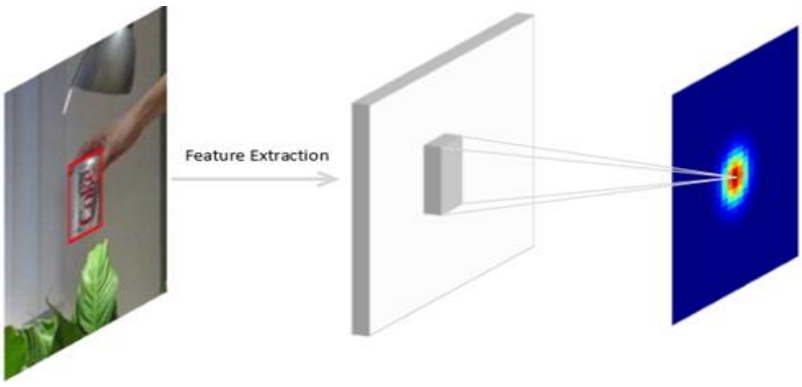



图 3-1 通过深度卷积方式进行回归

## 3.1 研究框架

判别相关滤波器（DCF）的成功是因为大量的样本被用来训练岭回归模型并预测物体的位置。但是 DCF 框架中依旧存在以下问题：（a）DCF 为了有效地解决回归问题，样本都是由一个搜索区域的循环移位产生的。然而，这些隐式构造的训练合成样本带来加速的同时，也会产生一些负面影响，削弱了基于 DCF 的跟踪器的鲁棒性。（b）训练模型时，样本中包含了太多的背景信息，这将会干扰到预测模型。同时，在构造深度 DCF 框架中，一个关键问题是超过 95%的训练样本是负样本，这将导致模型严重过拟合。（c）在传统 DCF 中没有考虑尺度变化问题。视频跟踪过程中，目标尺寸大小会随着距离镜头远近而不断变化，尺度保持不变会严重影响的性能。这三种负面效应限制了 DCF 的性能。

针对以上问题，在本章中，本文构建深度相关滤波器模型（DeepCF）。通过优化具有随机梯度下降（SGD）的单通道输出卷积层来解决回归问题：

1）在 DeepCF 中，提出了深度卷积回归框架。不同于传统的 DCF 模型封闭

式求解，只能解决浅层的模型，基于卷积回归的框架，通过优化一个具有梯度下降的特征通道输出卷积层来解决回归问题。不同于 DCF 中采用整体模板，通过深度卷积回归的方式可以通过小尺寸的卷积滤波器，滑窗采样真实场景中的样本，而非隐式构造的合成样本。

2）在 DeepCF 框架中为适应尺度的变化，融入尺度相关滤波器 DDST 作为独立目标的尺度估计组件。在连续的两帧中，目标位置的变化往往大于尺度的变化，因此，本文先采用深度卷积回归滤波器确定目标位置，在位置的基础上再使用尺度滤波器 DDST 确定当前目标的尺度。

## 3.2 基于卷积的回归

线性岭回归模型如何被用于视觉跟踪，如 CSK[26]，KCF[67]。给定一个带有回归标号的目标的初始图像

图 3-1 中通过卷积滤波器在输入图像（特征）上的扫窗实现回归器的训练样本采样，然后利用梯度下降法和反向传播技术对系数进行优化。与传统的判别相关滤波器相比，卷积回归采样样本是 “真实”场景的样本而非包涵大量背景的合成样本。在这一节中，本文介绍通过深度卷积网络的回归模型的方法。

### 3.2.1 解析式岭回归

首先，本文简单回顾线性岭回归模型如何被用于视觉跟踪，如 CSK[26]，KCF[67]。给定一个带有回归标号的目标的初始图像，提取大量的训练样本 $X \in \mathbb{R}^{m \times n}$，以及相应的回归标号 $\mathbf{y} \in \mathbb{R}^{m}$。这里 $m$ 是训练样本的个数，$n$ 是采样的维度。$X$ 的每一项 $x_i$，相应的回归目标为 $y_i$，即 $\mathbf{y}$ 的第 $i$ 个位置元素。然后，我们的目标是学习回归函数 $f(z)=w^T z$ 的系数 $w$，通过最小化下面的目标函数，

$$\arg\min_{w} \left\| X \cdot w - \mathbf{y} \right\|^2 + \lambda \left\| w \right\|^2 \tag{3-1}$$

这里，$\left\| \bullet \right\|^2$ 表示欧式范数正则项。由最小二乘上式损失函数表达式可以有封闭式解：

$$w = (X^T X + \lambda I)^{-1} X^T \mathbf{y} \tag{3-2}$$

然而，当 $m$ 和 $n$ 较大时（例如大于 1000，这在视觉跟踪中较为常见），采用上式

求解回归问题变得难以计算。这将限制岭回归在视觉跟踪中的应用。为了优化计算，在 DCF 中，通过从一个搜索区域中循环地转移样本来获得样本。然后将上式简化为基于循环技巧的高效计算。这一节，我们尝试用另一种方法来解决回归问题。

### 3.2.2 深度卷积滤波器回归

受卷积神经网络(CNN)的巨大成功的启发，本文提出学习一个梯度下降(GD)的回归模型。不同于 DCF 从单个搜索域中生成循环样本，而是通过滑动窗口来提取训练和预测样本。然后通过最后一层单通道输出的卷积层来计算这些样本的回归结果。与传统的相关滤波器使用整体模板不同的是，在深度特征上已具备较大的接受野，本文采用小的卷积模板，如3×3或5×5作为滤波器构建深度相关滤波器模型进行跟踪。

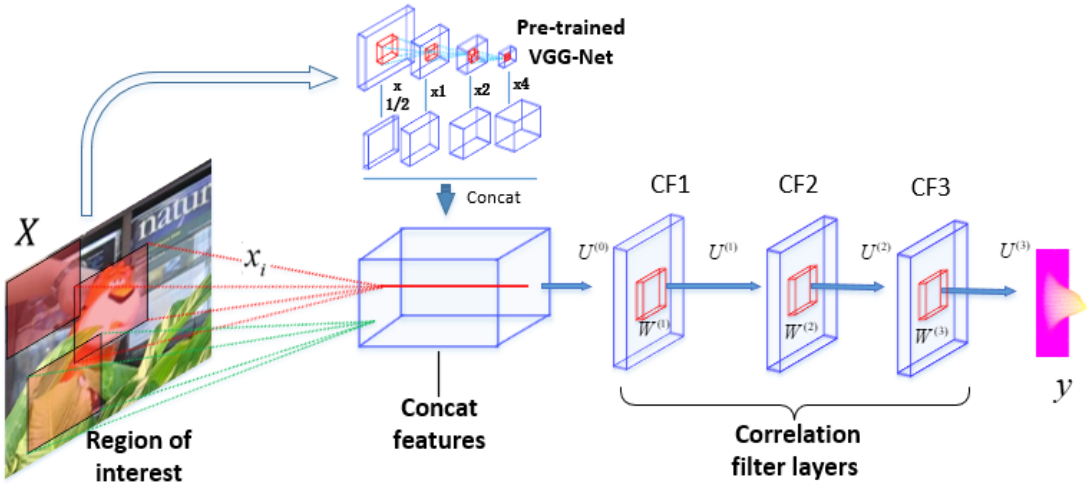



图 3-2 深度相关滤波器模型

本文采用两层的 CNN 滤波器作为相关滤波器回归最终的响应，本文的方法使用小尺寸的卷积模板学习物体的外观模型。由于在 VGG 网络中，浅层网络包含更多的空间纹理细节信息，深层网络包涵更多的语义类别信息 [55]，表征的特征信息丰富，因此本文使用预训练的深度 VGG-Net 中的不同层特征作为编码物体的外观特征。其中，网络中采用双线性插值将不同层由 pooling 操作导致不同特征层分辨率不同的特征插值到统一分辨率。

如上图 3-2 深度相关滤波器外观模型中，$X$ 表示输入的搜索域（Roi），经过

VGG-Net 预训练模型 $W_0$ 提取的特征 $U^{(0)} \in R^{C\times W\times H}$ 表示特征对应的尺寸为 $C\times W\times H$，这里 $W,H,C$ 代表宽，高和通道数。由于深度特征每一维特征向量 $U^{(0)}(x_{ij}) \in R^C$ 对应到搜索域的区域块 $x_{ij} \in X$，使用小尺寸的卷积滤波器扫窗采样代替 DCF 中将一系列采样样本通过预训练的网络提取特征的采样。$U^{(1)}$，$U^{(2)}$ 和 $U^{(3)}$ 分别表示三层深度相关滤波层的输出，并且 $W=\{W^{(1)},W^{(2)},W^{(3)}\}$ 和 $b=\{b^{(1)},b^{(2)},b^{(3)}\}$ 分别为 CF-CNN 的模型参数。相关滤波的两层输出可以表示为

$$\begin{cases} F^{(k)} = U^{(k-1)} * W^{(k)} + b^{(k)} \\ U^{(k)} = \kappa(F^{(k)}) \end{cases}, k = 1,2,3 \tag{3-3}$$

其中，$\kappa(x)=\max(0,x)$ 为 Relu 非线性激活函数，*表示卷积操作。在所有的层中本文没有使用 pool 层和全连接层，并且每一层通过 padding 技巧保持输出分辨率为 $W\times H$ 不变。$\mathbf{y}=\{y(i,j)\,|\,(i,j)\in\{0,1,...,W-1\}\times\{0,1,...,H-1\}\}$ 为高斯响应目标，其中 $y(i,j)=\exp(-(i-W/2)^2+(i-H/2)^2/2\sigma^2)$ 且 $\sigma$ 表示为高斯核半径。搜索域中的样本块 $x_{ij}\in X$ 对应回归目标为 $y(i,j)\in\mathbf{y}$，学习的 CF 模型参数为 $W=\{W^{(1)},W^{(2)},W^{(3)}\}$ 和 $b=\{b^{(1)},b^{(2)},b^{(3)}\}$，

$$\begin{aligned} J(W,b) &= minL_1 = min\left(\sum_{x_{ij}\in\chi}(U^{(3)}(x_{ij})-y(i,j))^2+\mu\,\|W\|^2\right) \\ &= min\left(\|U^{(3)}-\mathbf{y}\|_2^2+\mu\,\|W\|^2\right) \end{aligned} \tag{3-4}$$

其中，$\|W\|_2^2$ 为防止过拟合的正则项，$\lambda$ 为正则项的系数。卷积层的权值和偏置参数 $W$ 和 $b$，可以通过反向传播方程中定义的总损失来计算。利用梯度下降法迭代求解，得到近似最优解。

与 DCF 和简单的 CNN+DCF 相比，基于深度卷积的方法的优点有：（a）训练和检测阶段的样本都是真实场景的样本，这有助于提高性能和模型的鲁棒性。（b）整个图像搜索域中大量的负样本可以用来训练和更新回归模型，这将大大降低跟踪漂移的概率。（c）DeepCF 采用尺寸较小卷积模板作为相关滤波器，局

部编码方式引入更少的噪声信息，滤波器模型更加稳定。（d）标准的解析式相关滤波器只能构建单层的模型，而 DeepCF 中利用梯度下降方式求解的网络结构可以构建多层模型，更加灵活的构建跟踪模型。

## 3.3 基于深度相关滤波器的跟踪

基于卷积回归的视觉跟踪框架可以分解为训练、检测、更新三个阶段。本文分别用三段来解释，如图 3-4。

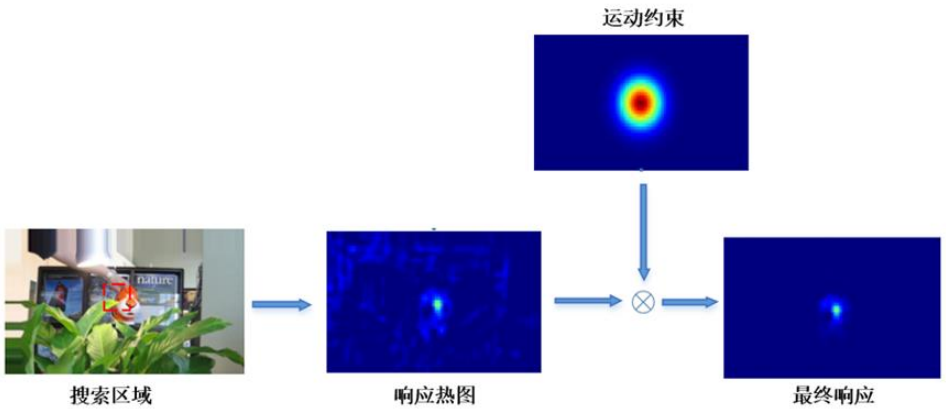



图 3-3 检测阶段流程

初始帧阶段：对于每个序列，本文首先基于初始的 Ground Truth 裁剪一块训练区域。由于跟踪序列的背景通常是静态的，因此被裁剪的训练区域应该比目标区域要大得多，以尽可能多地覆盖背景信息，从而降低了在后续帧中漂移的概率。然后，构建第 3.2.2 节中描述的卷积回归网络。其中，特征描述选取基于 ImageNet 预训练分 1000 类的 VGG-Net 特征。回归目标图可以用高斯函数来生成，其方差与物体的宽度和高度成正比。模型中的参数$W$ 和$b$，在初始化模型时，采用零均值高斯分布的随机初始化。随后的跟踪流程中，模型参数经过多次迭代，以更新系数，直到达到给定的损失阈值或有限的步骤。

检测阶段：在这个阶段，搜索区域的确定是裁剪以上一帧目标位置为中心的一块区域。然后，将搜索区域的提取特征传递到三层卷积回归网络中，输出为一个通道的响应热图。运动约束是利用高斯函数生成的，其变化与物体的大小成正比。最后的预测图是通过将运动约束和回归结果相乘来计算的。最终响应预测图最大值的位置即表示目标位置。检测过程如图 3-3 所示。对于尺度估计，本文采

用了 DSST 方法在目标位置确定后计算尺度。

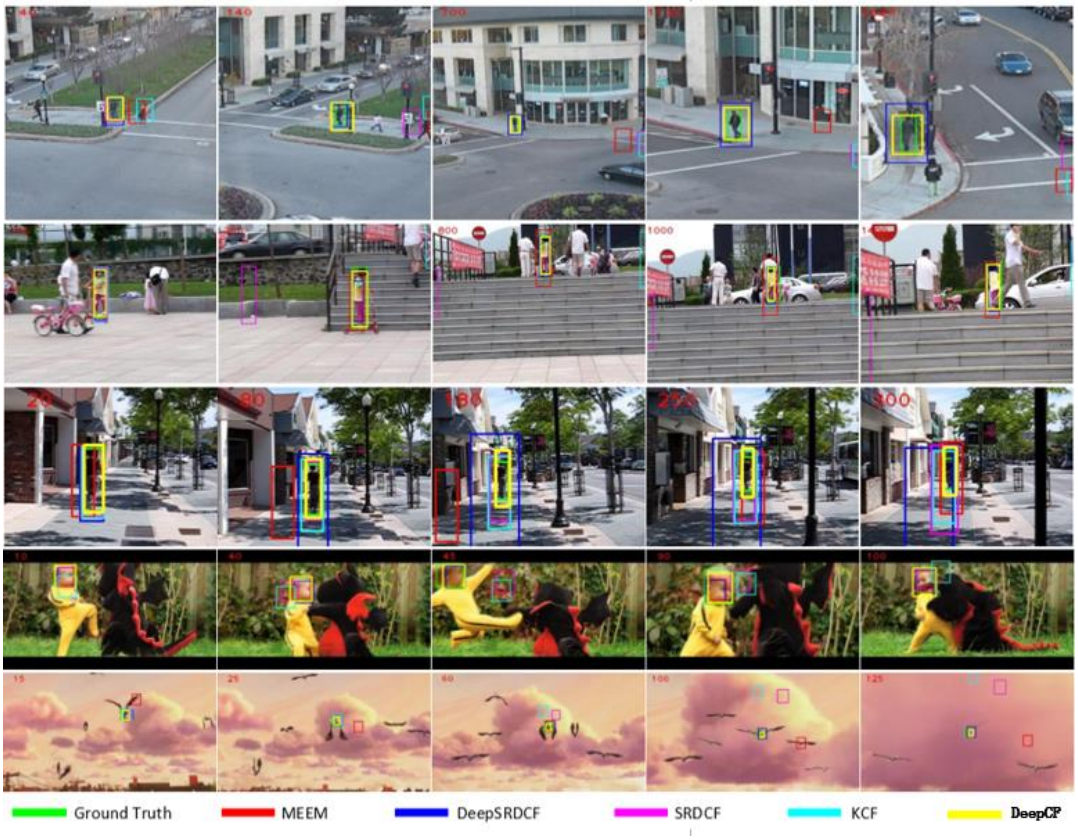



图 3-4 MEEM, DeepSRDCF, SRDCF, KCF 和 DeepCF 跟踪结果片段

更新阶段：为应对目标在运动过程中外观不断发生变化，跟踪过程中需要不断更新训练模型，以适应当前的目标外观变化和新的背景变化。同初始帧训练阶段，更新阶段的训练区域由当前帧检测结果作为依据生成。为了平滑地更新回归模型，将使用历史帧生成的训练的数据用于更新。在这个阶段，模型的参数系数通过随机梯度下降的进行更新。

## 3.4 本章小结

本章提出了深度相关滤波器（DeepCF）的框架。深度相关滤波器将传统判别的相关滤波器思想引入深度卷积神经网络中，代替传统封闭式求解方式回归，采用深度卷积滤波器的方式回归最终的响应。卷积滤波器采用的是小滤波器模板，且卷积扫窗操作等优点，缓解传统采用整体模板引入过多背景信息和边界效应等问题。深度相关滤波器较传统相关滤波器大幅度提升跟踪的性能。

# 第 4 章 自适应的深度相关滤波器

第三章中详细论述了传统判别相关滤波器到基于深度卷积的深度相关滤波器框架，以及在跟踪过程上的应用。本章以上一章研究工作为基础，着重研究了基于特征自适应在线学习的策略，提出了自适应的判别深度相关滤波器（adaDDCF）。

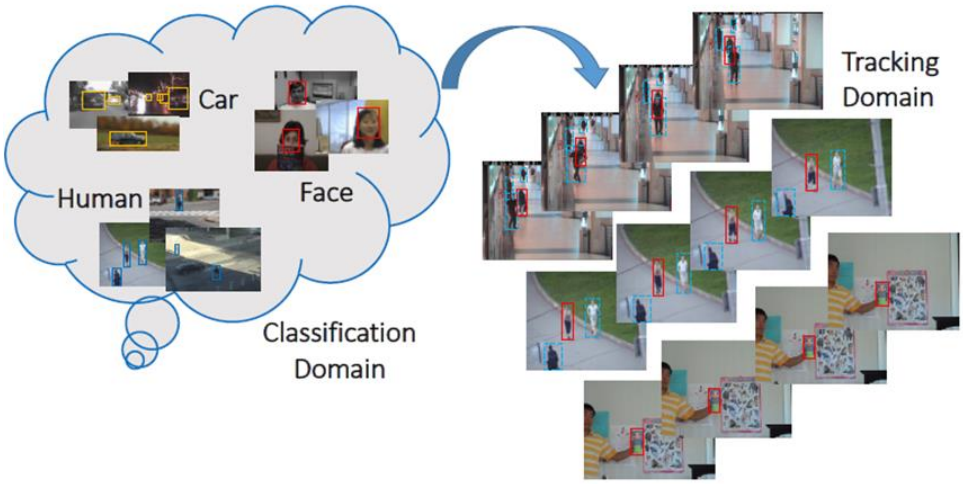



图 4-1 将分类域的特征用于跟踪域

## 4.1 研究框架

上一章中本文构建了 DeepCF 框架，一种卷积式的相关滤波器回归框架。DeepCF 利用梯度下降技术通过多层卷积层反向传播回归误差，对视觉跟踪的线性岭回归模型进行训练。与传统 DCF 相比，DeepCF 跟踪框架可以采样大量的“真实”样本，提高模型的判别能力。

在 DeepCF 框架中本文取代 DCF 传统的手工特征，使用了基于 ImageNet 预训练的分类特征。在使用分类特征存在两个缺点：（a）基于 ImageNet 大型数据集训练的分类网络只关注目标类别，而目标跟踪需要精确定位目标位置和大小，直接将分类域网络的特征用于跟踪域，其特征的判别力不够；（b）深度分类网络本身只关注特定的语义目标，然而，目标跟踪的对象并非一定是语义目标，因此，跟踪网络没有“看”过的目标时容易导致跟踪失败。

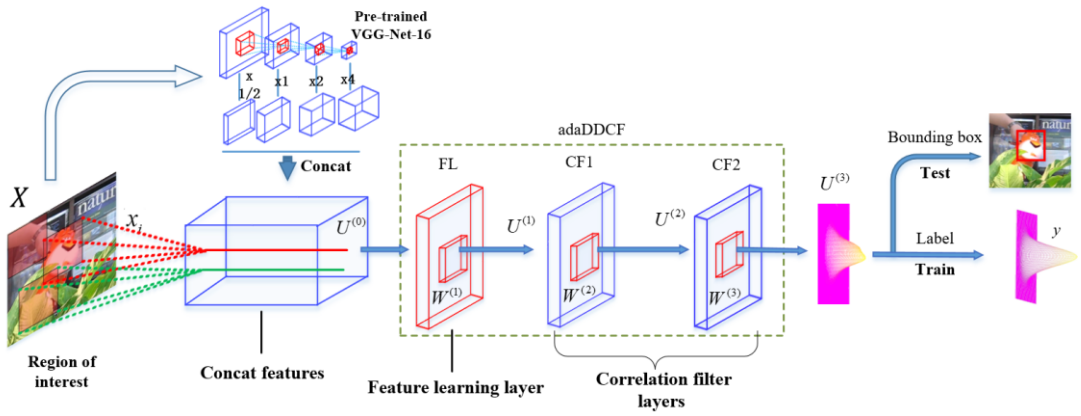



图 4-2 自适应的判别深度相关滤波器（adaDDCF）框架

在本章中本文提出了自适应的判别深度相关滤波器（adaDDCF）框架，在 DeepCF 基础上建立特征学习机制，将分类域的特征迁移学习至适合跟踪域的特征。由于基于分类数据集训练的特征网络对训练集中出现过的物体有很好的特征响应，但是在目标跟踪中被跟踪的物体极大可能是这个分类网络从未“看”到过的，因此需要将分类域的特征迁移学习致适合跟踪域的特征。基于此动机，本文采用一个特征学习策略在线动态的学习具有判别力的特征以适应跟踪域。

特征学习策略用于将分类任务预训练的深度分类域特征转换到适应跟踪问题的跟踪域特征。基于分类网络的训练数据有限，对目标跟踪问题中任意感兴趣的目标识别能力有限。文中采用一种全新的方式，通过引入和学习一个特殊的 Fisher 判别层来学习具有判别力的特征，学习判别力强的特征用于跟踪域。本文将搜索域内的前景目标和背景分成两类，Fisher 判别层通过利用在融合多层 CNN 特征上的 Fisher 判别准则在线训练模型，使得前景类和背景类同类的类内散度更小，不同类别的类间散度大，以此通过这种判别层使前景和背景特征更有区分性。

在 adaDDCF 中采用深度相关滤波器作为外观建模，并且耦合特征学习策略。深度相关滤波器建模外观在保证跟踪速度的前提下准确定位目标；耦合的特征学习策略能够学习具有判别力的特征，能够缓解模型漂移，带来跟踪性能的增益。

## 4.2 深度网络中的 Fisher 判别

本文的特征学习策略是被设计在深度相关滤波层前添加一层新的 Fisher 判别层，以预训练网络特征作为输入。基于 Fisher 判别准则的 Fisher 判别层强化学习

一个将前景和背景区分开判别能力强的特征，将搜索域分成目标和背景两类，使其两类特征有更小的类内散度和具有更大的类间散度。

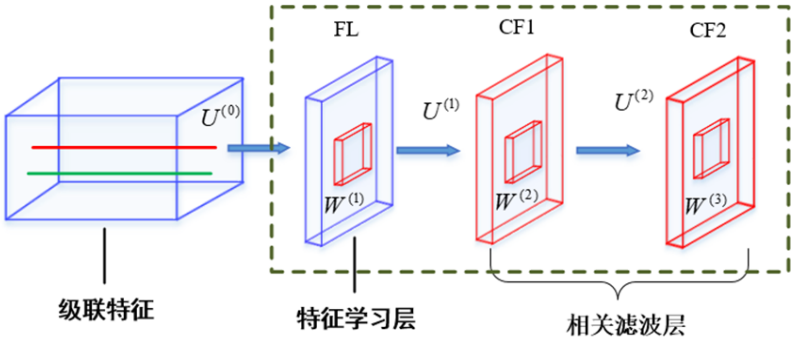



图 4-3 自适应的判别深度相关滤波器

网络输入图片搜索域为 $x$，$U^{(0)}$ 表示搜索域的预训练网络的特征输出。由于特征学习策略（FL）处于预训练特征之后，即 $U^{(1)} \in R^{K\times W\times H}$ 表示特征学习策略的输出。FL 特征学习策略层的模型参数分别为 $W^{(1)}$ 和 $b^{(1)}$。$U^{(1)}$ 的每个位置 $(i,j)$ 对应的向量为 $u_{ij}^{(1)} = U_{:,i,j}^{(1)} \in R^K$。在实验中，依据 Ground Truth 将搜索域 $X = \{X^+, X^-\}$ 分成前景 $X^+$ 和背景 $X^-$ 两类，对应特征学习策略层 $u_+^{(1)} = \{u_{ij}^{(1)}\}$ 和 $u_-^{(1)} = \{u_{ij}^{(1)}\}$ 分别为前景类和背景类特征。在本文采用的自适应的判别深度相关滤波器（DDCF）模型中除了要求模型最小化前景特征类间差（对应上一章节约束公式（3-4）），本文也要求特征学习策略模型对前景和背景有很强的判别能力。为此，结合上一节本文提出判别方程学习模型参数 $W = \{W^{(1)}, W^{(2)}, W^{(3)}\}$ 和 $b = \{b^{(1)}, b^{(2)}, b^{(3)}\}$，这里本文在深度相关滤波模型中耦合的加入特征学习策略，其损失函数可以更新为，

$$\begin{aligned} J(W,b) &= minL = min(L_1 + \lambda L_2) \\ &= min\left(\|U^{(3)} - \mathbf{y}\|_2^2 + \mu \|W\|^2 + \lambda(S_w - S_B)\right) \end{aligned} \tag{4-1}$$

这里 $\lambda$ 为平衡参数，控制两项的重要性。其中第一项和第二项为上一章节中方程（3-4），是一个卷积回归模型，最小化高斯响应目标与卷积相关响应的误差。

约束方程（4-1）中的第三项 $L_2$ 是一个 Fisher 判别正则约束项。依据 Fihser 判别准则输入样本 $X$ 分成前景 $X^+$ 和背景 $X^-$ 两类，最小化类内散度 $S_W(X)$，最大化

类间散度$S_B(X)$。$S_W(X)$和$S_B(X)$分别表示为，

$$\begin{aligned} S_w &= \sum_{u_{ij}^{(1)} \in u_+^{(l)}} (u_{ij}^{(1)} - \overline{u_+^{(l)}})(u_{ij}^{(1)} - \overline{u_+^{(l)}})^T + \sum_{u_{ij}^{(1)} \in u_-^{(l)}} (u_{ij}^{(1)} - \overline{u_-^{(l)}})(u_{ij}^{(1)} - \overline{u_-^{(l)}})^T, \\ S_B &= n^+(\overline{u_+^{(l)}} - \overline{u^{(1)}})(\overline{u_+^{(l)}} - \overline{u^{(1)}})^T + n^-(\overline{u_-^{(l)}} - \overline{u^{(1)}})(\overline{u_-^{(l)}} - \overline{u^{(1)}})^T, \end{aligned} \tag{4-2}$$

这里的$n^+$表示采样的前景类样本数量，对应为前景特征样本$u_+^{(1)}$的大小，同理$n^-$为背景类样本$u_-^{(1)}$数量。这里$\overline{u_+^{(l)}}$，$\overline{u_+^{(l)}}$和$\overline{u^{(1)}}$表示$u_+^{(1)}$，$u_-^{(1)}$和$U^{(1)}$特征样本的均值，可表示为，

$$\begin{aligned} \overline{u_+^{(l)}} = \frac{1}{n^+} \sum_{u_{ij}^{(1)} \in u_+^{(1)}} u_{ij}^{(1)}, \quad \overline{u_-^{(l)}} = \frac{1}{n^+} \sum_{u_{ij}^{(1)} \in u_-^{(1)}} u_{ij}^{(1)}, \\ \overline{u^{(1)}} = \frac{1}{n^+ + n^-} \sum_{x_i \in X} U^{(1)}, \end{aligned} \tag{4-3}$$

因此，上面的 Fisher 判别正则项$L_2$可以被定义如下，

$$L_2 = tr\left(S_W(X)\right) - tr\left(S_B(X)\right) \tag{4-4}$$

上述新的判别方程（4-1）不仅使用高斯软类别作为卷积回归标签，且学习一个新的更具有判别能力的 CNN 特征将前景和背景进一步分离，使得算法更加鲁棒。

### 4.3 可学习的 CNN 特征学习策略层

在跟踪中，为了在视频序列中跟踪任意被初始化的目标，本判别的相关滤波器模型中嵌入 Fisher 判别分析将适应于分类域的模型快速的迁移学习至适合跟踪域。在深度判别相关滤波器的前向过程中，FL 层以$U^{(l-1)}$作为输入生成$U^{(l)}$。在反向传播过程中，FL 层的梯度通过以下式子计算：

$$\begin{cases} W^{(l)} = W^{(l)} + \Delta W^{(l)} \\ \Delta W^{(l)} = -\eta\left(\dfrac{\partial L_1}{\partial W^{(l)}} + \lambda \dfrac{\partial L_2(U^{(l)})}{\partial W^{(l)}}\right) \end{cases}, 1 \le l \le 2, \tag{4-5}$$

式中$\eta$表示学习率，$\lambda$表示正则项的平衡参数。至于$L_2(U^{(l)})$这里的表示依赖于$U^{(l)}$，并且梯度$\partial L_2(U^{(l)})/\partial W^{(l)}$也依赖于$U^{(l)}$。梯度$\Delta W^{(l)}$依赖于$L_1$和$L_2$，使得模

型在外观建模的同时减少背景噪声。

此外，误差在网络层到的回传可以表示为，

$$\begin{aligned}\delta^{(l)} &= \frac{\partial[L_1 + L_2(U^{(l)})]}{\partial U^{(l)}} \\ &= \frac{\partial L_1}{\partial U^{(l+1)}}\frac{\partial(U^{(l+1)})}{\partial F^{(l+1)}}\frac{\partial(F^{(l+1)})}{\partial U^{(l)}} + \lambda\frac{\partial L_2(U^{(l)})}{\partial U^{(l)}} \\ &= \delta^{(l+1)}\kappa'(F^{(l+1)})\frac{\partial U^{(l)} * W^{(l+1)} + b^{(l+1)}}{\partial U^{(l)}} + \lambda L_2'(U^{(l)}) \\ &= \delta^{(l+1)}\kappa'(F^{(l+1)}) * W^{(l+1)} + \lambda L_2'(U^{(l)}), 1 \le l \le 2\end{aligned} \tag{4-6}$$

以上方程表示 FL 层的误差 $\delta^{(l)}$ 受 Fisher 判别误差和前一层的回归误差的影响。$\delta^{(l)}$ 的求解可进一步求得 $W^{(l)}$ 的梯度，

$$\begin{aligned}\frac{\partial L}{\partial W_{i,j}^{(l)}} &= \sum_{p,q}(\delta_j^{(l)})_{pq}(f_i)_{pq} \\ &= \sum_{p,q}[\delta_j^{(l+1)}\kappa'(F^{(l+1)}) * W_{j\cdot}^{(l+1)} + \lambda L_2'(U_j^{(l)})]_{pq}(f_i)_{pq}, 1 \le l \le 2\end{aligned} \tag{4-7}$$

同理 $b^{(l)}$ 的梯度可表示为，

$$\begin{aligned}\frac{\partial L}{\partial b_j^{(l)}} &= \sum_{p,q}(\delta_j^{(l)})_{pq} \\ &= \sum_{p,q}[\delta_j^{(l+1)}\kappa'(F^{(l+1)}) * W_{j\cdot}^{(l+1)} + \lambda L_2'(U_j^{(l)})]_{pq}, 1 \le l \le 2\end{aligned} \tag{4-8}$$

式中的中的 $i$ 和 $j$ 分别表示通道索引，$W_{j\cdot}^{(l+1)}$ 表示第 $l+1$ 层滤波器。$(f_i)_{pq}$ 表示 $U_i^{(l)}$ 在 $(p,q)$ 像素位置处的局部区域。

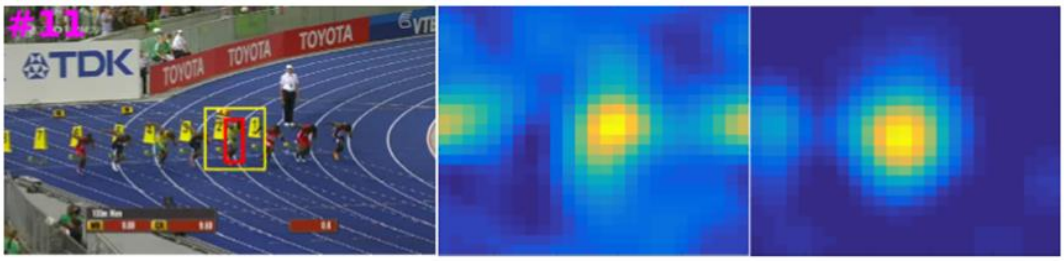



图 4-4 对比 DeepCF 与 adaDDCF 算法最后一层响应热图

从上面方程 4-7 和方程 4-8 可以看出，判别的特征学习和目标外观建模被

融合到一个端到端的跟踪框架中。在前向的传播过程中，判别的特征学习策略层 FL 学习了一个鲁棒的外观模型。

## 4.4 本章小结

在本章中本文提出了自适应的判别深度相关滤波器（adaDDCF）框架，在深度相关滤波器的基础上建立特征学习机制，将分类域的特征迁移学习致适合跟踪域的特征。采用深度相关滤波器作为外观建模，并且耦合特征学习策略。深度相关滤波器建模外观在保证跟踪速度的前提下准确定位目标；耦合的特征学习策略能够学习具有判别力的特征，能够缓解模型漂移，带来跟踪性能的增益。

# 第 5 章实验分析

第三、四章中分别详细论述了深度相关滤波器，和基于特征学习策略的自适应判别相关滤波器框架。本章节将实验验证本文算法的有效性。

## 5.1 实验简介

实验框架是基于 Matlab 和开源工具包 MatConvNet toolbox[72]实现，在主频 3.4GHZ CPU 和 Tesla K40 上运行速度为 9FPS。本算法采用深度神经网络随机梯度下降算法作为参数优化。

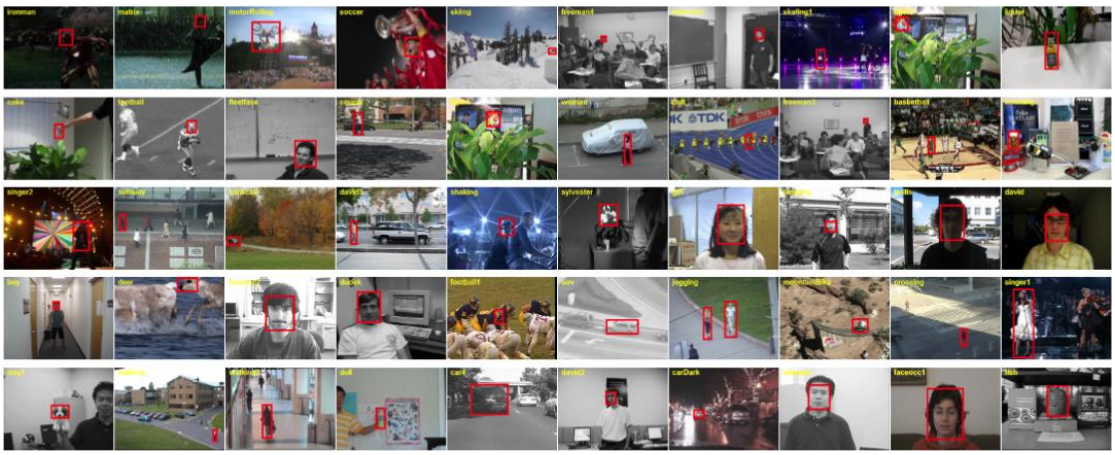


图 5-1 OTB2013 跟踪数据集

为了验证算法，选取目标跟踪公共数据集 OTB2013[70]，OTB2015[71]和 OTB50[71]作为算法评测数据集。其中 OTB2013 包涵 51 个视频序列，OTB2015 包涵 100 个视频序列，OTB50 选取 OTB2015 中更具有挑战的 50 个视频。所有的数据集包括 11 个属性，每个属性代表了视觉跟踪的一个挑战性问题。

实验中选取[71]中 29 个取得优异性能的算法和 14 个最近性能优异的算法作为对比实验。这些算法大致可以分成三类：1）基于传统手工特征的相关滤波器算法：SRDCFdecon[54]，KCF[67]，OCT-KCF[73]，SRDCF[59]，SAMF[74]和 DSST[64]。2）基于深度学习的相关滤波器跟踪算法：CFnet[49]，DeepSRDCF[61]和 HCF[60]。3）一些具有代表性的跟踪算法 CNN-SVM[63]，LCT[62]，

DLSSVM[57]，MEEM[65]，Staple[56]和 29 个[71]中代表性算法。

实验中所有算法都使用成功率曲线和精确度曲线作为性能评估。精确度曲线描述目标预测位置与真实位置之差小于一个给定阈值的曲线，成功率曲线描述了预测的目标框与真实目标框的重合率小于一定阈值的曲线。

1）准确率图：在跟踪精度评估中，一个被广泛使用的标准是中心位置误差，其被定义为跟踪目标的中心位置和手工标定的准确位置之间的平均欧氏距离。一个序列中所有帧的平均中心位置误差被用于慨括跟踪算法对该序列的总体性能。但是，当跟踪器丢失目标时，输出的跟踪位置是随机的，此时的平均误差值可能无法正确估量跟踪的性能。近年来，精确度图已经被用于测量跟踪的整体性能。精确度图能够显示出评估的位置在给定的准确值的阈值距离之内的帧数占总帧数的百分比。对于每个跟踪器具有代表性的精度评分，本文使用的分数阈值等于 20 个像素点。

2）成功率图：另一个评估标准是边界框的重叠率。假设跟踪的边界框为$\Phi_t$，准确的边界框是$\Phi_a$，重叠率被定义为$S=\left|\Phi_a \cap \Phi_t\right| / \left|\Phi_a \cup \Phi_t\right|$，其中$\cap$和$\cup$分别表示两个区域的交集和并集，$|.|$指其区域内的像素点个数。为了估量算法在一系列帧中的性能，本文计算重叠率 S 大于给定的阈值$\phi$的成功帧的数量。成功率图给出了此阈值从 0 到 1 变化时成功帧所占的比例。使用某一特定阈值（比如$\phi=0.5$）下的一个成功率来评估跟踪器可能并不公平或具有代表性。本文使用每一个成功率图的曲线下面积（AUC）作为替代，用于给跟踪算法进行排序。

此外实验中采用三个不同的评估实验评估：一次通过评估（OPE），时间鲁棒性评估（TRE）和空间鲁棒性评估（SRE）。

## 5.2 模型有效性实验

为证明判别的深度相关滤波器（adaDDCF）算法的有效性，本实验对比了 adaDDCF，HCF-var 算法和深度相关滤波器（Deep CF）算法。实验在 OTB2015 数据集上使用一次通过（OPE）进行评估，如图 5-3 模型有效性分析精确度曲线和成功率曲线。


图 5-2 adaDDCF、DeepCF 和 HCF 跟踪算法

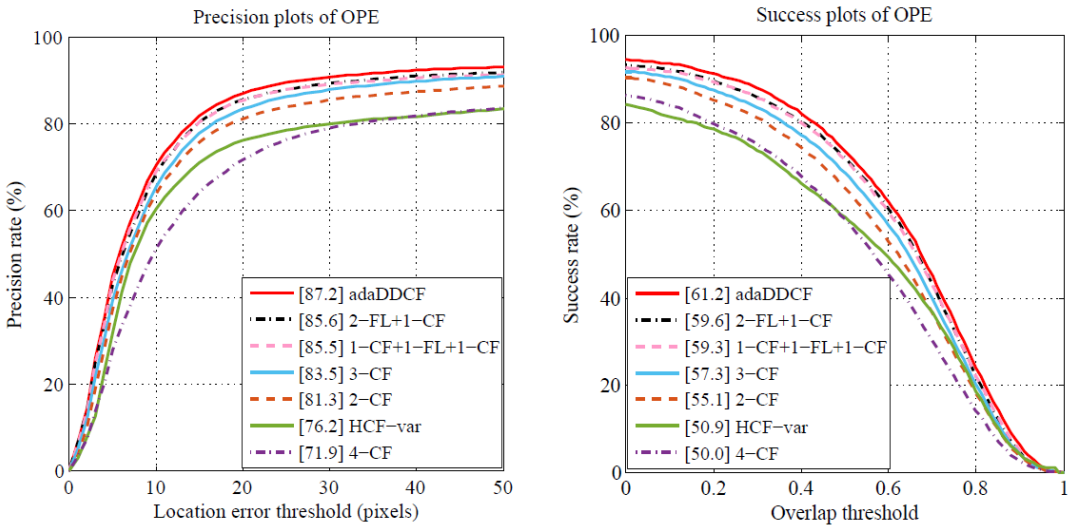



图 5-3 模型有效性分析

其中 HCF-var 算法表示原 HCF 算法的一个变形，将原有 HCF 算法特征描述替换成 adaDDCF 算法一致的基础特征描述。深度相关滤波器（Deep CF）表示未采用特征学习策略的深度相关滤波算法。图中 $k-CF$ 表示一个 $k$ 层的深度相关滤波器，$1-CF+1-FL+1-CF$ 表示一个两层的深度相关滤波器中间层加入特征学习策略，$2-FL+1-CF$ 表示两层特征学习策略层后面连接一层卷积滤波层。本文中的 adaDDCF 采用了 $1-FL+2-CF$ 的结构，即一层特征学习策略后面连接两层相关滤波层。

上图 5-3 中，首先，基于深度学习的 $k-CF$ 的跟踪器相对于基于传统手工特

征相关滤波器性能上有显著提升，并且$3-CF$相对于$1-CF$和$2-CF$有更优异的性能，但是由于深度卷积层数的增加，带来的参数增加会引入过拟合问题，因此$4-CF$跟踪器相对与前面三种跟踪器算法性能上有较大下降。其次，对比加入特征学习的深度相关滤波器（adaDDCF，$2-FL+1-CF$ 和$1-CF+1-FL+1-CF$）与$k-CF$，证明特征学习策略极大提升跟踪性能。

| Tracker | Precision | | Success Rate | |
|---|---|---|---|---|
| | thr=20 | Average | thr=20 | Average |
| adaDDCF | 87.2 | 80.0 | 73.5 | 61.2 |
| correlation filter with deep CNN | | | | |
| DeepSRDCF | 85.0 | 79.4 | 76.4 | 63.5 |
| HCF | 83.7 | 76.7 | 71.8 | 56.2 |
| CFnet | 77.7 | 71.7 | 73.2 | 58.8 |
| correlation filter with hand-crafted features | | | | |
| SRDCFdecon | 82.4 | 77.2 | 75.8 | 62.7 |
| SRDCF | 78.9 | 74.0 | 72.0 | 59.7 |
| OCT-KCF | 75.7 | 70.2 | 58.9 | 52.3 |
| SAMF | 74.4 | 70.0 | 66.7 | 53.5 |
| DSST | 69.3 | 65.5 | 59.6 | 52.0 |
| KCF | 69.2 | 64.7 | 54.4 | 47.6 |
| 5 recently representative trackers, and the top 1 tracker from Wu et al. | | | | |
| CNN-SVM | 81.4 | 74.7 | 64.6 | 55.4 |
| Staple | 78.4 | 73.1 | 70.1 | 57.8 |
| LCT | 78.9 | 70.0 | 69.3 | 59.8 |
| MEEM | 78.1 | 71.3 | 61.3 | 53.0 |
| DLSSVM | 76.7 | 71.1 | 61.6 | 54.1 |
| Struck | 63.9 | 59.7 | 51.7 | 46.0 |

表 5-1 OTB2015 数据集对比实验

## 5.3 OTB2015 数据集实验评估

OTB100 数据集包涵了 100 个视频序列，是 OTB 系列最全面的数据集。为验证算法的有效性，图 5-4 和表 5-1 中展示了本文的 adaDDCF 算法对比 14 个最近最具代表性的算法 29 个跟踪算法[71]（这里只列出前 15 个算法）。

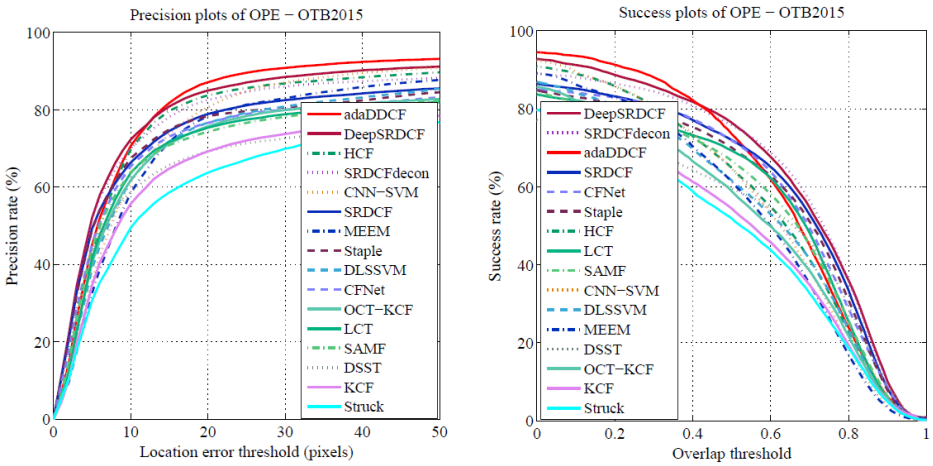



图 5-4 OTB2015 数据集对比实验

由于特征学习策略的有效的抑制预训练网络在跟踪中出现的过拟合现象，adaDDCF 在精确度曲线上达到一个最优的性能，在成功率曲线上排名第三。其中，成功率略低于 DeepSRDCF 和 SRDCFdecon 两个算法，但是本文的 adaDDCF 算法在运行速度上分别是 DeepSRDCF 和 SRDCFdecon 两个算法的 9 倍和 3 倍。综上，在 OTB2015 上评测的实验表明，相对于一些代表性跟踪算法本文中的 adaDDCF 算法具有显著优势。

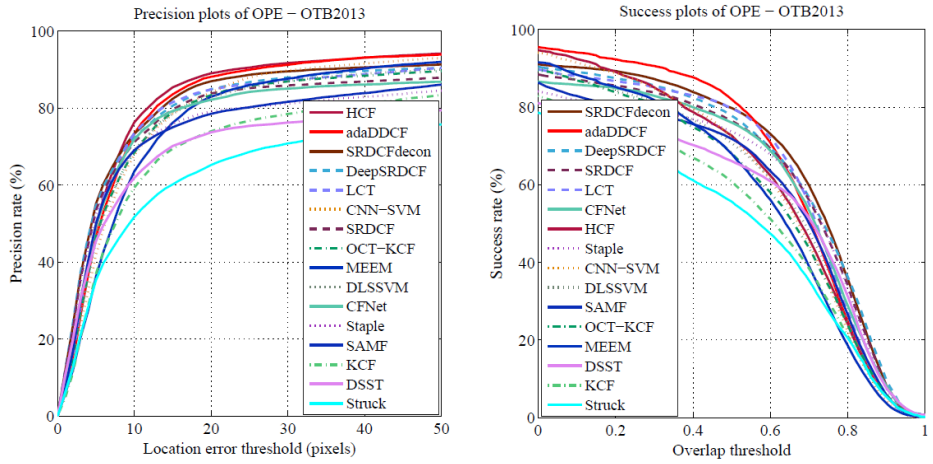



图 5-5 OTB2013 数据集对比实验

## 5.4 OTB2013 数据集实验评估

图 5-5 中详细展示了 OTB2015 数据集 50 个视频上的性能实验，实验包涵了 adaDDCF 算法对比 14 个最近最具代表性的算法和 29 个跟踪算法[45]（这里只列出前 15 个算法）。

| Tracker | Precision | | Success Rate | |
|---|---|---|---|---|
| | thr=20 | Average | thr=20 | Average |
| adaDDCF | 88.2 | 80.9 | 82.0 | 64.3 |
| correlation filter with deep CNN | | | | |
| HCF | 89.1 | 81.8 | 72.7 | 60.5 |
| DeepSRDCF | 84.9 | 79.4 | 77.8 | 64.1 |
| CFnet | 82.2 | 76.5 | 76.0 | 61.0 |
| correlation filter with hand-crafted features | | | | |
| SRDCFdecon | 87.0 | 80.6 | 79.8 | 65.3 |
| SRDCF | 83.8 | 77.6 | 76.5 | 62.5 |
| OCT-KCF | 83.4 | 77.0 | 68.2 | 57.4 |
| SAMF | 78.5 | 73.9 | 71.8 | 58.0 |
| DSST | 73.7 | 68.7 | 66.1 | 55.3 |
| KCF | 74.0 | 68.9 | 60.9 | 51.4 |
| 5 recently representative trackers, and the top 1 tracker from Wu et al.[71] | | | | |
| CNN-SVM | 84.1 | 78.2 | 72.3 | 59.2 |
| LCT | 83.3 | 77.2 | 79.8 | 62.3 |
| MEEM | 83.0 | 75.9 | 68.0 | 56.6 |
| DLSSVM | 82.9 | 77.0 | 70.8 | 58.9 |
| Staple | 79.2 | 74.0 | 73.7 | 60.0 |
| Struck | 65.4 | 60.8 | 55.6 | 47.1 |

表 5-2 OTB2013 数据集对比实验

本文中的 adaDDCF 跟踪算法在 OTB2013 上有较为显著的优势。依据表 5-2 中， adaDDCF 算法在精确度和成功率上分别比 Struct（[71]中 29 个代表性算法中性能最优算法）算法高出 22.8%和 26.4%。相比基于传统手工特征的相关滤波器，本文的跟踪算法 adaDDCF 分别实现了 4.4%和 5.5%的性能增益。虽然 adaDDCF 算法的精确度略低于 HCF 算法， 但是 adaDDCF 算法在成功率上大大超越 HCF 算法 9.3%。总的来说，在 51 个有挑战性的视频（OTB2013）上的实验结果评测表明 adaDDCF 跟踪器算法相对于最先进的跟踪算法具有更优的表现。

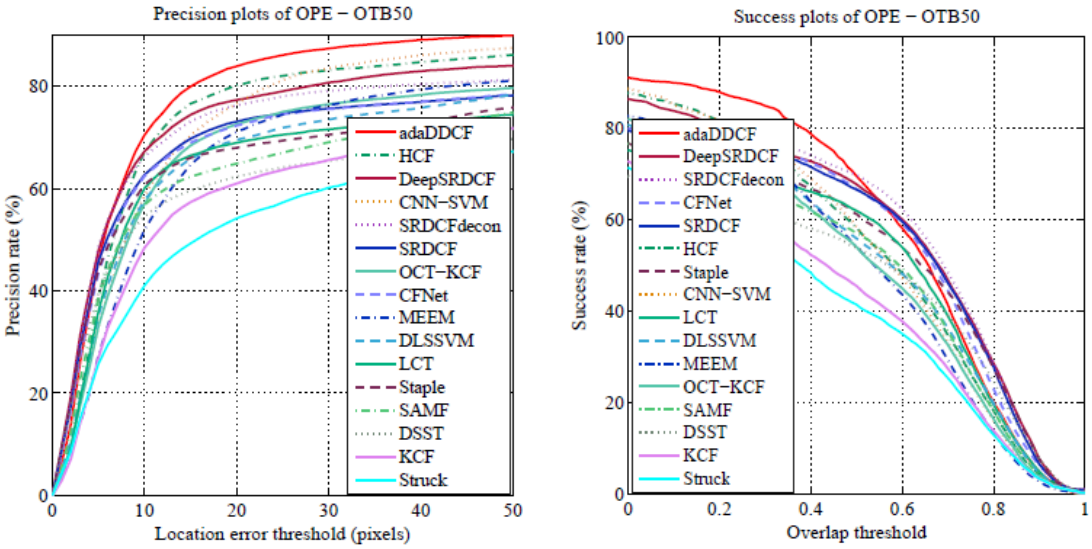



图 5-6 OTB50 数据集对比实验

| Tracker | Precision | | Success Rate | |
|---|---|---|---|---|
| | thr=20 | Average | thr=20 | Average |
| adaDDCF | 83.9 | 77.5 | 69.5 | 57.7 |
| correlation filter with deep CNN | | | | |
| HCF | 80.1 | 73.3 | 58.3 | 51.3 |
| DeepSRDCF | 77.3 | 72.9 | 67.7 | 55.9 |
| CFnet | 72.3 | 67.8 | 65.6 | 53.8 |
| correlation filter with hand-crafted features | | | | |
| SRDCFdecon | 76.3 | 71.4 | 69.8 | 55.9 |
| SRDCF | 73.1 | 68.1 | 66.6 | 53.8 |
| OCT-KCF | 72.6 | 66.5 | 54.0 | 47.2 |
| SAMF | 64.9 | 62.1 | 57.1 | 47.0 |
| DSST | 62.4 | 59.3 | 53.8 | 46.2 |
| KCF | 61.0 | 57.3 | 45.5 | 40.2 |
| 6 recently representative trackers, and the top 1 tracker from Wu et al.[71] | | | | |
| CNN-SVM | 76.6 | 71.4 | 58.4 | 51.1 |
| MEEM | 71.1 | 65.2 | 54.2 | 47.2 |
| DLSSVM | 69.5 | 64.8 | 55.9 | 48.7 |
| LCT | 69.0 | 63.7 | 61.9 | 49.3 |
| Staple | 68.0 | 64.6 | 61.2 | 51.2 |
| Struck | 57.3 | 52.1 | 41.9 | 38.4 |

表 5-3 OTB50 数据集对比实验

## 5.5 OTB50 数据集实验评估

OTB50 数据集是源于 OTB2015 数据集更具有挑战性的子集，包涵 50 个视频序列。如图 11，本文在 OTB50 上评测了 adaDDCF 算法对比 14 个最近最具代表性的算法和[71]中 29 个跟踪算法（这里只列出前 15 个算法）。

依据表格 5-3 中详细性能指标，本文的 adaDDCF 跟踪算法在精确度和成功率上分别优于第二种最佳方法的是 4.2%和 1.8% 。在 OTB50 数据集测试中所展示的结果，本文中的 adaDDCF 跟踪算法在最具挑战的 50 个视频上取得具有竞争力的性能。

## 5.6 鲁棒性实验

评估跟踪器的传统方式是，根据第一帧中的准确位置进行初始化，然后在一个测试序列中运行算法，最后得出平均精确度或成功率的结果报告。本文把这种方法成为一次通过的评估（OPE）。然而跟踪器可能对初始化非常敏感，并且在不同的初试帧给予不同的初始化会使其性能变得更差或更好。因此，本文提出两种方式来评估跟踪器对初始化的鲁棒性，即在时间上（即在不同帧开始跟踪）和空间上（即以不同的边界框开始跟踪）扰乱初始化。为了进一步验证算法的鲁棒性，采用除了一次通过评估（OPE）的时间鲁棒性评估（TRE）和空间鲁棒性（SRE）评估。

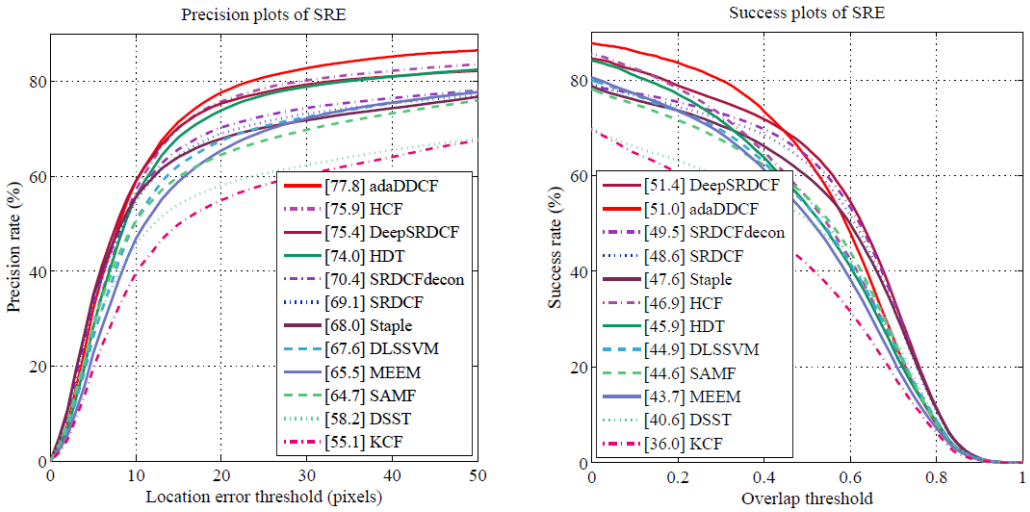



图 5-7 空间鲁棒性评估（SRE）

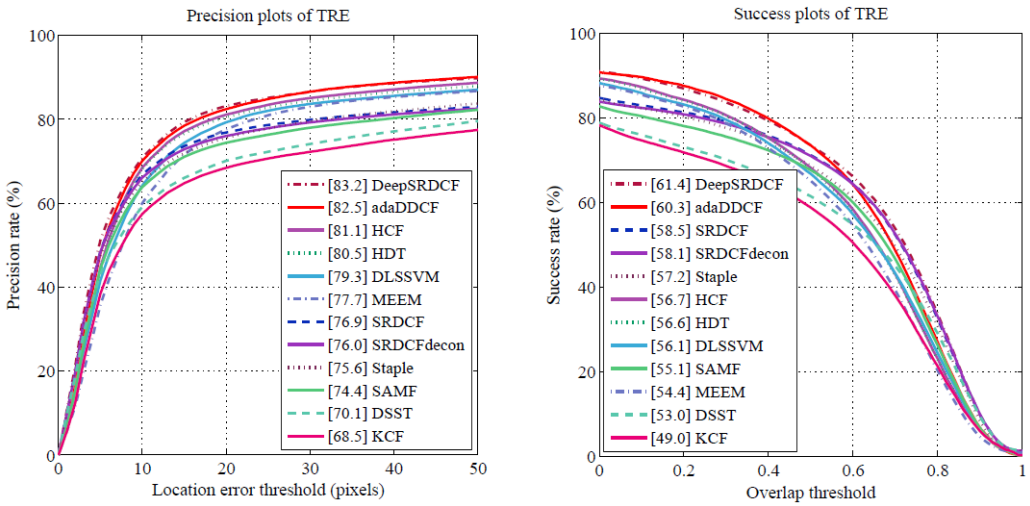



图 5-8 时间鲁棒性评估（TRE）

所提出的测试场景大部分就存在于现实世界中实际应用的场合，而跟踪器通常通过目标检测器来初始化，检测器在位置或尺寸方面可能会给跟踪器引入初始化误差。另外，在不同时刻的实例中，检测器可能被用于重新初始化跟踪器。通过研究跟踪器在不同鲁棒性评估中的特点，本文可以对跟踪算法进行更为深入的理解和分析。

1）时间鲁棒性评估（TRE）：给定一个标记了目标准确边界框的初试帧，跟踪器被初始化并运行直到序列结束，即整个序列的一部分。跟踪器会在每一个序列的片段上进行评估且整体的统计数据也会被记录下来。

2）空间鲁棒性评估（SRE）：本文在第一帧中通过移动或缩放准确的 ground truth 来抽取初始化的边界框。在这里，本文使用 8 种空间位置上的偏移，包括 4 种中心偏移和 4 种角偏移，以及 4 种尺度变化。偏移量为目标尺寸的 10%，尺度比例变化可取准确值的 0.8、0.9、1.1 和 1.2。因此，针对 SRE 本文对每个跟踪器评估 12 次。

本文选取了排名前 11 的跟踪算法与本文的 adaDDCF 算法进行实验对比。adaDDCF 算法在空间鲁棒性评估上取得平均（精确度和成功率）第一的性能，在时间鲁棒性上取得仅次于 DeepSRDCF 算法的性能。

## 5.7 属性实验

由于诸多因素会影响跟踪的性能，评估跟踪算法是困难的。为了更好地评

估和分析跟踪方法的优点和缺点，本文用 11 种属性标注所有序列来进行分类，这些属性列在表 5-4 中。

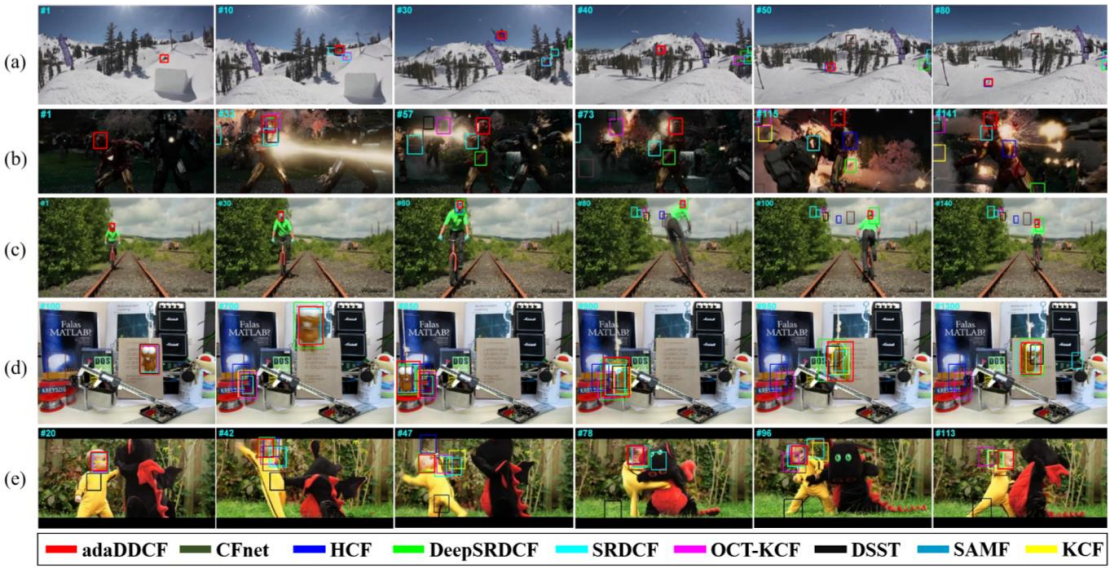



图 5-9 不同属性下算法跟踪序列效果

| | |
|---|---|
| IV | 光照变化 -目标区域内的光照剧烈变化 |
| SV | 尺度变化 -第一帧中和当前帧中的边界框尺寸之比的范围超过 [1/t, t]，其中 t >1 (t=2) |
| OCC | 遮挡 -目标被部分或全部遮挡 |
| DEF | 形变 -非刚体目标发生形变 |
| MB | 运动模糊 -目标或摄像机的运动导致的目标区域变模糊 |
| FM | 快速运动 - Ground Truth 的运动大于 20 个像素点 |
| IPR | 平面内旋转 -目标在图像平面内发生旋转 |
| OPR | 平面外旋转 -目标在图像平面外发生旋转 |
| OV | 超出视野 -目标的一部分离开视野 |
| BC | 背景杂乱 -目标附近的背景具有和目标类似的颜色或纹理 |
| LR | 低分辨率 - ground-truth 边界框内的像素点个数少于 400 个 |

表 5-4 目标跟踪数据集中 11 种属性

在基于特性的实验分析中，选取了 8 个相关滤波器相关的跟踪算法进行对比实验，分别是：HCF，SRDCFdecon，DeepSRDCF，KCF，OCT-KCF，SRDCF ，SAMF 和 DSST。本文的属性实验基于 OTB50 进行评估。本文选择了 5 个主要的具有挑战性的属性，提供了以下的点具体分析：

1) 变形（DEF）：在跟踪过程中，目标在局部变形(非刚性)，导致目标外观发生变化，导致跟踪不稳定性。图 5-10 显示了非刚性变形情况下的跟踪性能，表明具有自适应特征学习策略和深度相关滤波器外观模型的 adaDDCF 跟踪算法优于其同系列算法。图 5-9(a)所示的定性跟踪结果表明，刚性变形发生的情况。adaDDCF 跟踪器在跟踪其出现明显的非刚性变形的“滑雪人”时，跟踪效果良好。但是，追踪过程中没有特征学习策略的追踪器 CFnet、DeepSRDCF、SRDCF、OCT-KCF、DDST、SAMP 和 KCF 都出现了跟踪失败。

2) 平面外的旋转：由于视角的变化和目标的运动，在连续的视频序列中，目标的出现经常受到平面外的旋转影响。图 5-11 给出了 adaDDCF 跟踪器的跟踪性能，表明 adaDDCF 的性能优于 8 个同系列算法。在图 5-9 (b)中，定性跟踪的结果来自几个代表性的框架，其中“钢铁侠”的出现在跟踪过程中，由于外平面旋转而发生了剧烈的变化，采用特征学习策略，有效地跟踪了“钢铁侠”，从而在这类具有挑战性的框架中获得了较好的跟踪结果。

3) 低分辨率：图像中像素有限导致目标缺乏对外观表示的详细信息，从而导致检测和跟踪的难度大。图 5-12 展示提 adaDDCF 算法的跟踪性能和低分辨率下的列外 8 个跟踪器算法的性能。adaDDCF 跟踪器在精确度方面表现最好。然而，adaDDCF 在成功率评测下获得了第 4 的性能，这是由于 adaDDCF 的适度变化适应能里相对较弱。图 5-1 0(c)显示了几个代表性帧的跟踪结果展示，尽管“骑车人”的头部的像素相对较少， adaDDCF 跟踪器在这个序列中依旧表现良好。

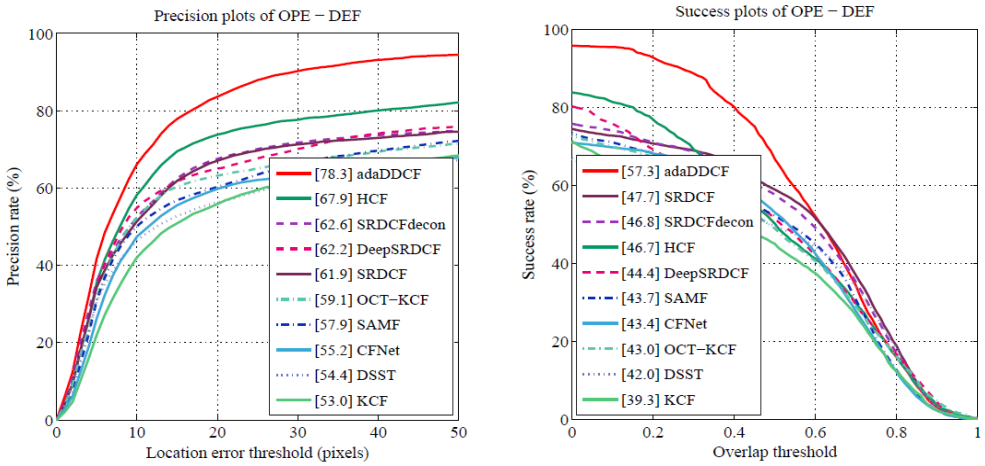



图 5-10 形变属性下各算法性能（OTB50）

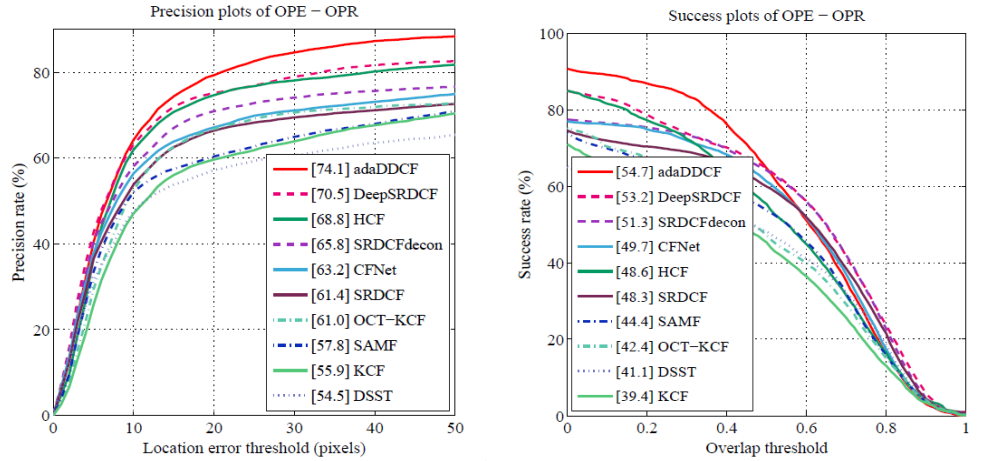



图 5-11 平面外旋转属性下各算法性能（OTB50）

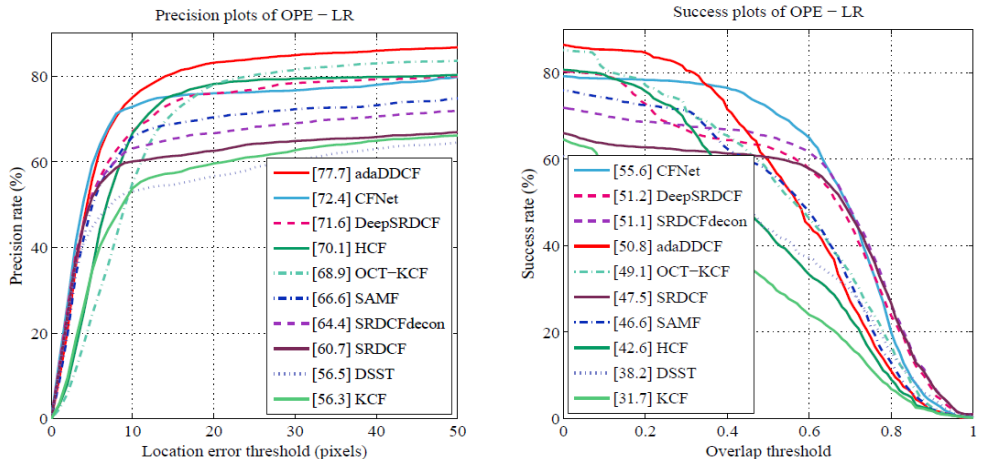



图 5-12 低分辨率属性下各算法性能（OTB50）

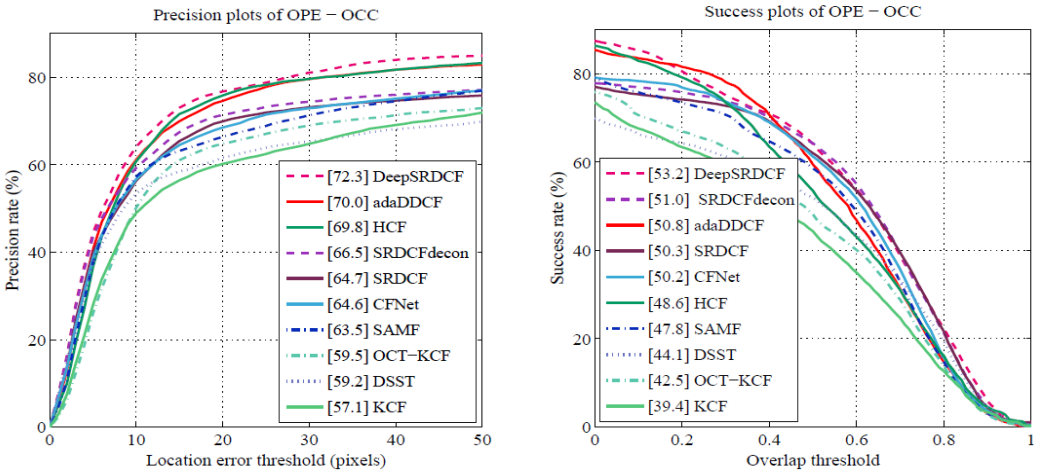



图 5-13 遮挡属性下各算法性能（OTB50）

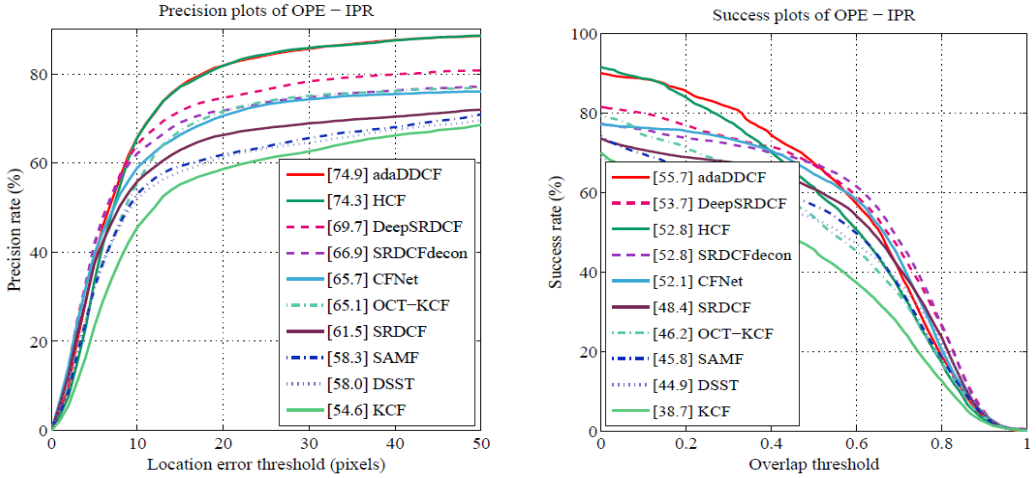



图 5-14 平面外旋转属性下各算法性能（OTB50）

4) 遮挡：在跟踪时，目标往往会被其他物体遮挡住，导致出现部分或完全的消失，因此遮挡是导致跟踪失败的主要因素之一。图 5-13 详细说明了 adaDDCF 的跟踪性能和在遮挡情况下的同类跟踪器的跟踪性能。图 5-10(d)所示的跟踪结果，其中玩具熊在凌乱的桌子上移动，有时被其他物体遮挡住。adaDDCF 通过对其特点的改进和改进，对目标的外观变化进行了适应性的学习，并获得了与同类跟踪算法相比更稳定的跟踪性能。

5) 平面内旋转：在跟踪过程中，目标的运动往往导致出现平面内旋转，在这种情况下，很难估计目标和背景之间的边界。这是影响目标定位精度的因素之一。视觉跟踪器的跟踪性能如图 5-14 所示，在这种情况下，所提出的 adaDDCF 跟踪器在精度和成功率方面的表现都优于排名第二的跟踪算法。图 5-

10(e)所示的定性跟踪结果表明，尽管在跟踪过程中，在摄像机平面中跟踪的“小孩”的脸部在平面内旋转，adaDDCF 跟踪器在视频跟踪中表现优秀。

## 5.8 本章小结

本章对第三章中的深度相关滤波器外观模型和第四章中自适应的深度相关滤波器进行详细的实验验证。第一节，介绍了实验的相关配置和数据集。第二节，对比了深度相关滤波器、自适应的深度相关滤波器以及相关框架的对比实验，以证明本文模型的有效性，并且分析自适应的深度相关滤波器的对比优劣。第三、四、五章，分别在 OTB2015、OTB2013 和 OTB50 数据集上进行了本文的自适应的深度相关滤波器与当前最优秀的跟踪算法的对比实验，并且分析自适应的深度相关滤波器的优势。第六章，为测试自适应的深度相关滤波器的鲁棒性，在 OTB50 上进行了时间鲁棒性实验和空间鲁棒性实验。第七章，对自适应的深度相关滤波器进行了 11 中具有挑战性的属性实验。

# 第 6 章 总结与展望

目标跟踪可应用于智能视频监控、车辆辅助驾驶以及智能机器人等方面，具有重要的研究价值和实际应用价值。本文主要研究深度相关滤波器，以及基于特征学习的深度相关滤波器—adaDDCF 在目标跟踪中的应用。将传统的相关滤波器迁移至深度学习领域，建立深度相关滤波器的跟踪器，并且在此基础上耦合特征学习策略优化跟踪的性能。

## 6.1 本文工作总结

为解决传统相关滤波器由循环矩阵、整体滤波器和只能封闭求解带来的局限，本文中提出了深度相关滤波器模型，并且应用于目标跟踪。对于深度相关滤波器中采用固定分类域的特征不能很好适应跟踪域特征的问题，提出了基于特征学习的自适应深度相关滤波器模型。总结如下：

(1) 深度相关滤波器（DeepCF）。深度相关滤波器将传统判别的相关滤波器思想引入深度卷积神经网络中，代替传统封闭式求解方式回归，采用深度卷积滤波器的方式回归最终的响应。卷积滤波器采用的是小滤波器模板，且卷积扫窗操作等优点，缓解传统采用整体模板引入过多背景信息和边界效应等问题。深度相关滤波器较传统相关滤波器大幅度提升跟踪的性能。

(2) 自适应的判别深度相关滤波器（adaDDCF）。在深度相关滤波器的基础上建立特征学习机制，将分类域的特征迁移学习致适合跟踪域的特征。采用深度相关滤波器作为外观建模，并且耦合特征学习策略。深度相关滤波器建模外观在保证跟踪速度的前提下准确定位目标；耦合的特征学习策略能够学习具有判别力的特征，能够缓解模型漂移，带来跟踪性能的增益。

## 6.2 未来工作展望

随着智能监控、无人驾驶以及智能机器人和复杂的交通场景这些应用的驱动，对跟踪的性能和效率提出更高的要求。深度学习在跟踪上的应用，大幅度提升跟

踪性能，但是由于深度网络的复杂度较高，导致基于深度的跟踪算法效率普遍较低。其次，目标跟踪中一大难题是对遮挡目标的跟踪，由于遮挡的问题常常导致跟踪失败。由此，未来可以从如下几个方面进行后续研究，完善视觉目标跟踪：

（1）优化网络提高跟踪效率：大多基于深度学习的跟踪算法和本文的自适应的深度相关滤波器算法相对传统方法已经大大提高了跟踪的性能。然而，在追求不断提高的追踪性能方面，其特征速度和实时性能力逐渐淡化。

（2）对语义目标的跟踪：深度学习在跟踪上的应用，较传统方法，已经大幅度提升跟踪性能。但是在面临实际出现遮挡、外形非刚性变化大的场景时，依旧会导致跟踪失败。因此需要网络“预学习”被跟踪物体的多样性外观，这样可解决在跟踪时因外观遮挡或是变化较大带来的问题，模型鲁棒性更强。

# 参考文献


[1] Li X, Hu W, Shen C, et al. A survey of appearance models in visual object tracking[J]. Acm Transactions on Intelligent Systems & Technology, 2013, 4(4):1-48.

[2] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[C]. International Joint Conference on Artificial Intelligence. Morgan Kaufmann Publishers Inc. 1981:674-679.

[3] Matthews I, Ishikawa T, Baker S. The Template Update Problem[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2004, 26(6):810-815.

[4] Alt N, Hinterstoisser S, Navab N. Rapid selection of reliable templates for visual tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010:1355-1362.

[5] Adam A, Rivlin E, Shimshoni I. Robust Fragments-based Tracking using the Integral Histogram[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006:798-805.

[6] Hager G D, Belhumeur P N. Efficient Region Tracking With Parametric Models of Geometry and Illumination[M]. IEEE Computer Society, 1998.

[7] Ross D A, Lim J, Lin R S, et al. Incremental Learning for Robust Visual Tracking[J]. International Journal of Computer Vision, 2008, 77(1-3):125-141.

[8] Mei X, Ling H. Robust Visual Tracking using l(1) Minimization[C]. IEEE International Conference on Computer Vision, 2009:1436-1443.

[9] Mei X, Ling H. Robust Visual Tracking and Vehicle Classification via Sparse Representation. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(11):2259–2272.

[10] Ahuja N. Robust visual tracking via multi-task sparse learning[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:2042-2049.

[11] Zhong W, Lu H, Yang M H. Robust object tracking via sparse collaborative appearance model.[J]. IEEE Transactions on Image Processing, 2014, 23(5):2356.

[12] Lu H, Jia X, Yang M H. Visual tracking via adaptive structural local sparse appearance model[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1822-1829.

[13] Ji H, Ling H, Wu Y, et al. Real time robust L1 tracker using accelerated proximal gradient approach[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1830-1837.

[14] Comaniciu D, Ramesh V, Meer P. Kernel-Based Object Tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2003, 25(5):564-575.

[15] Pérez P, Hue C, Vermaak J, et al. Color-Based Probabilistic Tracking[J]. European Conference

on Computer Vision, 2002, I:661-675.

[16] Collins R T. Mean-shift Blob Tracking through Scale Space[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003:234.

[17] Isard M, Blake A. CONDENSATION—Conditional Density Propagation for Visual Tracking[J]. International Journal of Computer Vision, 1998, 29(1):5-28.

[18] Birchfield S T, Rangarajan S. Spatiograms versus histograms for region-based tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005:1158-1163 vol. 2.

[19] He S, Yang Q, Lau R W H, et al. Visual Tracking via Locality Sensitive Histograms[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013:2427-2434.

[20] Dalal N, Triggs B. Histograms of Oriented Gradients for Human Detection[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005:886-893.

[21] Tang F, Brennan S, Zhao Q, et al. Co-Tracking Using Semi-Supervised Support Vector Machines[J]. IEEE International Conference on Computer Vision, 2007, 1:1-8.

[22] Porikli F. Integral histogram: a fast way to extract histograms in Cartesian spaces[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005:829-836 vol. 1.

[23] H. Grabner, M. Grabner, and H. Bischof. Real-Time Tracking via On-line Boosting[C]. British Machine Vision Conference, 2006.

[24] Babenko B, Yang M H, Belongie S. Robust Object Tracking with Online Multiple Instance Learning[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(8):1619-32.

[25] Zhang K, Zhang L, Yang M H. Real-Time Compressive Tracking[C]. European Conference on Computer Vision, 2012:864-877.

[26] Rui C, Martins P, Batista J. Exploiting the circulant structure of tracking-by-detection with kernels[C]. European Conference on Computer Vision, 2012:702-715.

[27] T. Ojala, M. Pietik ̈ainen, and T. M ̈aenp ̈ a ̈a. Multiresolution Grayscale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis & Machine Intelligence, 24(7):971–987, 2002.

[28] Viola P, Jones M. Robust real-time face detection[J]. International Journal of Computer Vision, 2004, 57(2):137-154.

[29] Hinton G. E., Salakhutdinov R. R.. Reducing the dimensionality of data with neural networks[J]. Science, 2006, 313(5786): 504–507.

[30] Krizhevsky A., Sutskever I., Hinton G. E.. Imagenet classification with deep convolutional neural networks[C]. Proceedings of Advances in Neural Information Processing Systems, 2012: 1106–1114.

[31] Liwicki S, Zafeiriou S P, Pantic M. Online Kernel Slow Feature Analysis for Temporal Video Segmentation and Tracking[J]. IEEE Transactions on Image Processing, 2015, 24(10):2955-2970.

[32] Nayak N M, Zhu Y, Roy Chowdhury A K. Hierarchical Graphical Models for Simultaneous Tracking and Recognition in Wide-Area Scenes[J]. IEEE Transactions on Image Processing, 2015, 24(7):2025-36.

[33] Godec M, Roth P M, Bischof H. Hough-based tracking of non-rigid objects[C]. International Conference on Computer Vision. IEEE Computer Society, 2011:81-88.

[34] Lu H, Jia X, Yang M H. Visual tracking via adaptive structural local sparse appearance model[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1822-1829.

[35] Fan J, Wu Y, Dai S. Discriminative Spatial Attention for Robust Tracking[C]. European Conference on Computer Vision, 2010:480-493.

[36] Avidan S. Support Vector Tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003:I-184-I-191 vol.1.

[37] Avidan S. Ensemble Tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(2):261-271.

[38] Hare S, Saffari A, Torr P H S. Struck: Structured output tracking with kernels[C]. IEEE International Conference on Computer Vision, 2012:263-270.

[39] Bai Y. Robust tracking via weakly supervised ranking SVM[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1854-1861.

[40] Grabner H. Real-Time Tracking via On-line Boosting[J]. British Machine Vision Conference, 2006.

[41] Babenko B, Yang M H, Belongie S. Robust Object Tracking with Online Multiple Instance Learning[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(8):1619-32.

[42] Avidan S. Support Vector Tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003:I-184-I-191 vol.1.

[43] Collins R T, Liu Y, Leordeanu M. Online Selection of Discriminative Tracking Features[M]. IEEE Computer Society, 2005.

[44] Avidan S. Ensemble Tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007, 29(2):261-271.

[45] Babenko B, Yang M H, Belongie S. Robust Object Tracking with Online Multiple Instance Learning[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2011, 33(8):1619-32.

[46] Zhong W. Robust object tracking via sparsity-based collaborative model[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2012:1838-1845.

[47] Yu Q, Dinh T B. Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers[C]. European Conference on Computer Vision, 2008:678-691.

[48] Danelljan M, Bhat G, Khan F S, et al. ECO: Efficient Convolution Operators for Tracking[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016:6931-6939.

[49] Valmadre J, Bertinetto L, Henriques J, et al. End-to-End Representation Learning for Correlation Filter Based Tracking[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2017:5000-5008.

[50] Held D, Thrun S, Savarese S. Learning to Track at 100 FPS with Deep Regression Networks[M]. European Conference on Computer Vision, 2016:749-765.

[51] 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, 2016:472-488.

[52] Nam H, Han B. Learning Multi-domain Convolutional Neural Networks for Visual Tracking[J]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015:4293-4302.

[53] Wang L, Ouyang W, Wang X, et al. STCT: Sequentially Training Convolutional Networks for Visual Tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016:1373-1381.

[54] Danelljan M, Häger G, Khan F S, et al. Adaptive Decontamination of the Training Set: A Unified Formulation for Discriminative Visual Tracking[C], IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016:1430-1438.

[55] Qi Y, Zhang S, Qin L, et al. Hedged Deep Tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016:4303-4311.

[56] Bertinetto L, Valmadre J, Golodetz S, et al. Staple: Complementary Learners for Real-Time Tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016:1401-1409.

[57] Ning J, Yang J, Jiang S, et al. Object Tracking via Dual Linear Structured SVM and Explicit Feature Map[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016:4266-4274.

[58] Wang L, Ouyang W, Wang X, et al. Visual Tracking with Fully Convolutional Networks[C]. IEEE International Conference on Computer Vision, 2015:3119-3127.

[59] Danelljan M, Hager G, Khan F S, et al. Learning Spatially Regularized Correlation Filters for Visual Tracking[C]. IEEE International Conference on Computer Vision, 2016:4310-4318.

[60] Ma C, Huang J B, Yang X, et al. Hierarchical Convolutional Features for Visual Tracking[C]. IEEE International Conference on Computer Vision, 2016:3074-3082.

[61] Danelljan M, Häger G, Khan F S, et al. Convolutional Features for Correlation Filter Based Visual Tracking[C]. IEEE International Conference on Computer Vision Workshop, 2016:621-629.

[62] Ma C, Yang X, Zhang C, et al. Long-term correlation tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2015:5388-5396.

[63] Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network[C]. International Conference on International Conference on Machine Learning, 2015:597-606.

[64] Danelljan M, Häger G, Khan F S. Accurate scale estimation for robust visual tracking[C]. British Machine Vision Conference. 2014:65.1-65.11.

[65] Zhang J, Ma S, Sclaroff S. MEEM: Robust Tracking via Multiple Experts Using Entropy Minimization[C]. European Conference on Computer Vision, 2014:188-203.

[66] Danelljan M, Hager G, Khan F S, et al. Discriminative Scale Space Tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(8):1561-1575.

[67] Henriques J F, Rui C, Martins P, et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 37(3):583-596.

[68] Nam H, Baek M, Han B. Modeling and Propagating CNNs in a Tree Structure for Visual Tracking[J]. arXiv, 2016.

[69] Ning G, Zhang Z, Huang C, et al. Spatially supervised recurrent convolutional neural networks for visual object tracking[C]. IEEE International Symposium on Circuits and Systems, 2017:1-4.

[70] Wu Y, Lim J, Yang M H. Online Object Tracking: A Benchmark[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2013:2411-2418.

[71] Wu Y, Lim J, Yang M H. Object Tracking Benchmark[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(9):1834-1848.

[72] Vedaldi A, Lenc K. MatConvNet:Convolutional Neural Networks for MATLAB[C]. ACM International Conference on Multimedia, 2015:689-692.

[73] Zhang B, Li Z, Cao X, et al. Output constraint transfer for kernelized correlation filter in tracking[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 47(4): 693-703.

[74] Li Y, Zhu J. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration[C]. European Conference on Computer Vision, 2014:254-265.

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

[76] Danelljan M, Khan F S, Felsberg M, et al. Adaptive Color Attributes for Real-Time Visual

Tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2014:1090-1097.

[77] Fan J, Xu W, Wu Y, et al. Human tracking using convolutional neural networks[J]. IEEE Transactions on Neural Networks, 2010, 21(10): 1610-1623.

[78] Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[C]. Advances in Neural Information Processing Systems, 2013: 809-817.

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

[80] Fan H, Ling H. Sanet: Structure-aware network for visual tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2017: 2217-2224.

[81] Tao R, Gavves E, Smeulders A W M. Siamese instance search for tracking[C]. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016: 1420-1429.

[82] 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.

[83] Ren, Shaoqing, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence 2015:1137-1149.

[84] He, Kaiming, et al. Mask R-CNN[C]. IEEE International Conference on Computer Vision, 2017:2980-2988.

[85] Li, Hanxi, Y. Li, and F. Porikli. DeepTrack: Learning Discriminative Feature Representations Online for Robust Visual Tracking. IEEE Transactions on Image Processing, 2015:1834-1848.

[86] Cai B, Xu X, Xing X, et al. BIT: Biologically Inspired Tracker.[J]. IEEE Transactions on Image Processing, 2016, 25(3):1327-1339.

# 致 谢

在中国科学院大学攻读硕士期间，我在科研上经历了不少的挫折，也得到了很大的锻炼和收获。在毕业论文即将完成之际，我由衷地感谢给我帮助的老师、同学和家人。

本课题的研究工作是在韩振军副教授，焦建彬教授以及叶齐祥教授的悉心指导下完成的。感谢韩振军老师在科研的过程中对我的的指导和鼓励，在我学习和科研遇到困难，难以前行的时候，鼓励我，并为我提供思路。感谢焦建彬教授在我攻读硕士学位期间从学习和生活各个方面给予的无微不至的关怀与指导，以及在我论文的撰写和修改中倾注的心血。感谢叶齐祥老师在论文写作的过程中给予的指导和帮助以及生活上给予的关心。恩师们在科研上精益求精，在学术上认真严谨，他们的科学精神令人敬佩；恩师们在生活上关心爱护学生，和蔼可亲、平易近人，他们春风化雨的教诲让人感动。

感谢各位师兄、师姐，在我研究生的三年中给予的学习上的耐心引导与生活中的种种帮助。感谢我的同届好友以及师弟师妹们，三年的科研生活中大家相互帮助、献计献策、相互提点，一起渡过了三年快乐的日子。这些快乐的时光在我记忆中永远不会褪色。

感谢参加开题及中期评阅的各位老师和专家们，他们丰富的经验和无私的工作对论文方向和研究进度的把握和指点给整个研究工作带来了许多帮助。

王 攀

2018 年 5 月

# 作者简历及攻读学位期间发表的学术论文与研究成果

## 作者简历：

2009 年 09 月——2013 年 07 月，在中国石油大学信息与控制工程学院获得学士学位

2013 年 8 月——2014 年 10 月，在中国船舶重工-武汉船用机械技术中心任职电气工程师

2015 年 09 月——2018 年 06 月，在中国科学院大学电子电气与通信工程学院攻读硕士学位

## 在审论文：

[1] **Pan Wang,** Zhaoju Li, Shan Gao, Baochang Zhang, Qixiang Ye, Zhenjun Han*. Adaptive Discriminative Deep Correlation Filter for Robust Visual Tracking[J]. *submitted to Elsevier Neurocomputing*. (SCI)