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： 基于卷积神经网络的目标跟踪技术研究

专　　业： 模式识别与智能系统

研究方向： 计算机视觉

姓　　名： 李振邦

学　　号： 201818014628082

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：2020 年 12 月 8 日

目　　录

基于卷积神经网络的目标跟踪技术研究

一、研究背景与意义

目标跟踪是计算机视觉领域中最重要和最具挑战性的研究课题之一。目标跟踪的核心是估计图像序列的每帧中目标的运动状态。目标跟踪是计算机视觉领域的中层部分，为目标的行为理解提供了基础，因此具有非常重要的理论研究价值。同时，它具有广泛的实际应用，包括视频监控，交通流量监控，视频压缩和人机交互等。例如，目标跟踪已成功应用于监控居民区，停车场和银行中的人类活动（例如 W4 系统[1]和 VSAM 项目[2]）。在交通运输领域，目标跟踪也被广泛用于交通流量监控[3]，行人计数[4]等任务。

由于目标跟踪的理论价值与应用价值，众多科研机构和公司都投入到这项研究中。然而，目标跟踪领域存在很多理论和技术问题有待解决，如运动模糊、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。近年来深度学习的突破为解决目标跟踪中的一系列问题带来了可能。

深度学习是基于人工神经网络的机器学习方法。在过去的十年中，深度学习技术得到了飞速发展，已成功应用于计算机视觉，语音识别，自然语言处理，音频识别，社交网络过滤，机器翻译，生物信息学，药物设计等领域。如何利用深度学习方法，尤其是深度卷积神经网络解决跟踪过程中遇到的复杂问题，具有较大的研究价值和研究空间。

二、国内外研究现状

本章主要介绍基于深度学习的目标跟踪的研究现状，将基于深度学习的跟踪器划分为以下几种类别：基于深度学习的相关滤波跟踪器，基于生成对抗网络的跟踪器，基于图卷积网络的跟踪器，基于循环神经网络的跟踪器，基于孪生网络的跟踪器，基于强化学习的跟踪器，基于无监督学习的跟踪器，基于注意力机制的跟踪器和基于串并联或级联结构的跟踪器。

a) 基于深度特征的相关滤波跟踪器

相关滤波跟踪器是传统跟踪器中的代表之一。传统的相关滤波跟踪器通常采用手工设计的特征或底层特征，这限制了相关滤波器的潜力。随着深度学习时代的到来，有很多相关滤波跟踪器尝试使用深度特征代替底层特征，并取得了性能的提升。[5]利用从深度卷积神经网络中提取的特征进行相关滤波的学习，来提高跟踪精度和鲁棒性。作者利用最后一个卷积层的输出对目标的语义信息进行编码，使得跟踪器对于目标的表观变化具有鲁棒性。但是，由于空间分辨率太粗糙，无法精确定位目标。相反，较浅的卷积层提供了更精确的定位。因此作者在每个卷积层上自适应学习相关滤波器，以对目标表观进行编码。在[6]中，作者在常规相关滤波跟踪器框架的基础上，提出了一种用于训练连续卷积滤波器的算法。作者采用隐式插值模型来解决连续空间域中的学习问题。该模型可实现对于多种分辨率的特征图的有效集成。此外，该方法能够进行亚像素定位，有利于提高跟踪的精确性。

b) 基于生成对抗网络的跟踪器

生成对抗网络（GAN）可以通过 CNN 从随机噪声生成逼真的图像。 生成对抗网络包含两个子网，一个充当生成器，另一个充当判别器。 生成器旨在合成图像以欺骗判别器，而判别器则试图正确区分真实图像和生成器合成的图像。通过相互竞争来同时训练生成器和鉴别器。对抗学习的优势在于，所训练的生成器可以生成与训练样本相似的图像统计信息，从而使判别器无法区分。生成对抗网络的进步吸引了包括目标跟踪在内的各种计算机视觉应用的关注。在[7]中，作者利用生成对抗网络产生的样本辅助跟踪器的学习。文章指出，由于以下问题，现有视觉跟踪器的性能可能会受到限制：i）采用密集采样策略生成的正样本会降低样本的多样性；ii）即使收集到大规模的训练数据集，具有挑战性的训练数据也是有限的。作者提出了 VITAL 算法来通过对抗学习解决这两个问题。为了增加正样本，作者使用一个生成网络随机生成模板，这些模板用于自适应过滤输入特征以捕获各种表观变化。通过对抗学习获得的模板，可以提供最鲁棒的目标特征。此外，为了解决类别不平衡的问题，作者提出了一个高阶成本敏感损失，从而有助于训练分类网络。在[8]中，作者通过对抗生成学习产生难例正样本进行跟踪。具体来说，作者假设目标都位于流形上，因此，引入正样本生成网络（PSGN），通过遍历已构建的目标流形来采样大

量训练数据。生成的各种目标图像可以丰富训练数据集并增强目标跟踪器的鲁棒性。为了使跟踪器对遮挡更加鲁棒，作者提出了一个变换网络，该网络可以生成用于跟踪算法的难例样本。

c) 基于图卷积网络的跟踪器

图卷积神经网络（Graph Convolutional Network）是一种能对图数据进行深度学习的方法。在目标跟踪中，图卷积网络用于捕获目标样本的结构特征。在[9]中，作者指出时空信息可以用于增强目标表示，并且上下文信息对于目标的定位很重要。为了全面利用历史目标样本的时空结构并从上下文信息中受益，作者提出了一种用于高性能视觉跟踪的新型图卷积跟踪（GCT）方法。具体而言，GCT 将两种类型的图卷积网络（GCN）合并到用于目标表观建模的孪生框架中。作者采用时空 GCN 来建模历史目标样本的结构化表示。而上下文 GCN 被设计为利用当前帧的上下文来学习用于目标定位的自适应特征。在[10]中，作者同样使用 GCN 模块来学习目标跟踪的结构特征。首先，作者利用双路径网络提取异构特征。然后，作者采用 GCN 模块来构建具有结构化信息的要素。

d) 基于循环神经网络的跟踪器

循环神经网络（Recurrent Neural Network, RNN）是一类以序列（sequence）数据为输入，在序列的演进方向进行递归（recursion）且所有节点（循环单元）按链式连接的神经网络。RNN 在建模序列数据方面引起了越来越多的关注。这些应用程序涵盖了多语言机器翻译[11]，动作识别[12,13]，场景标记[14,15]，语音识别[16]等。最近，传统的 RNN 被归纳为更复杂的结构模型，例如二维 RNN [17,18]，多维 RNN [19,20,21]，树 RNN [22,23]等。在目标跟踪中，可利用 RNN 建模目标的复杂远程依赖关系。RTT[24]尝试识别并利用那些对整个跟踪过程有益的可靠部分。为了解决遮挡并发现可靠的组件，RTT 中使用了多方向递归神经网络，通过从多个方向遍历候选空间区域来捕获远程上下文线索。从 RNN 生成的置信度图用于抑制背景噪声，同时充分利用来自可靠部分的信息，来自适应地区分判别相关滤波器的学习。在[25]中，作者提出了一种能够将时间信息整合到模型中的实时目标跟踪器。该跟踪器不是专注于有限的一组目标或在测试时训练一个模型来跟踪特定的实例，而是在大量不同的目标上预先训练一个通

用跟踪器，并进行实时的在线更新。

e) 基于孪生网络的跟踪器

孪生网络是一种用于度量学习的有监督模型。孪生网络具有两个参数共享的子网络，可以学习两幅输入图像之间的特征相似性。由于优越的性能，基于孪生网络的跟踪器已经成为当前目标跟踪领域的主流。在 SiamFC[26]中证明了使用孪生网络解决跟踪问题的有效性。具体来说，作者训练了一个孪生网络以在较大的搜索图像中定位模板图像。利用互相关操作以滑动窗口的方式获得目标位置的响应图，从而对目标进行实时定位。在 SiamRPN[27]中，跟踪器由用于特征提取的孪生网络和包括分类分支和回归分支的 region proposal 子网络组成。受益于跟踪器的改进，传统的多尺度测试和在线微调可以被丢弃。

f) 基于强化学习的跟踪器

强化学习（RL）的目标是学习一种通过最大化未来累积奖励来决定动作序列的策略。在[28]中，作者提出了一种新颖的跟踪器，该跟踪器顺序执行通过深度强化学习而学到的动作来进行控制。与使用深层网络的现有跟踪器相比，所提出的跟踪器旨在实现轻量级计算以及令人满意的跟踪精度。控制动作的深层网络使用各种训练序列进行了预训练，并在跟踪过程中进行了微调，以在线适应目标和背景变化。在[29]中，作者将跟踪形式化为部分可观察的决策过程（POMDP）来学习最佳决策策略。作者使用深度强化学习算法学习策略，这些算法仅在运动轨迹出现问题时才需要监督（奖励信号）。作者证明稀疏的奖励有利于快速地对海量数据集进行训练。

g) 基于无监督学习的跟踪器

无监督学习（unsupervised learning）是机器学习的一种方法，没有给定事先标记过的训练样本，自动对输入的数据进行分类或聚类。常见的目标跟踪方法往往需要以监督方式进行训练，需要大量带注释的真实标签。手动注释往往是昂贵且费时的，而大量的未标记视频可在 Internet 上轻松获得。通过无监督学习，可以利用未标记的视频序列进行视觉跟踪。在[30]中，作者通过使用辅助自然图像，离线训练堆叠式去噪自动编码器，以学习对变化更鲁棒的通用图像特征。然后，将知识从离线培训转移到在线跟踪过程。在线跟踪网络由训练过的自动编码器（作为特征提取器）和一个附加的分类层构成。特征提取器

和分类器都可以进行在线更新以适应运动目标的表观变化。在[31]中，作者提出了一种无监督的视觉跟踪方法。与使用大量有标签数据进行监督学习的现有方法不同，作者提出的 CNN 模型是在无监督的大规模无标签视频上进行训练的。作者的动机是，强大的跟踪器在前向和后向预测中均应有效（即，跟踪器可以在连续帧中向前定位目标并回溯到其在第一帧中的初始位置）。作者在一个孪生相关滤波网络上构建跟踪框架，该网络使用未标记的原始视频进行训练。同时，文中提出了一种多帧验证方法和一种成本敏感的损失函数，以促进无监督学习。

h) 基于注意力机制的跟踪器

注意力机制首先用于神经科学领域[32]。它们已经扩展到其他领域，例如图像分类[33,34,35]，姿态估计[36]等。在目标跟踪领域中，注意力机制有利于使网络的学习关注更有效的信息。 RASNet 模型[37]在孪生跟踪框架内重新构造了相关过滤器，并引入了各种注意机制来适应模型而无需在线更新模型。通过利用离线训练的通用注意力，目标自适应的残差注意力以及通道特征注意力，RASNet 不仅减轻了深度网络训练中的过拟合问题，而且具有更强的判别能力和适应性，从而提高了跟踪的性能。文中提出的深度架构是端到端训练的，充分利用了丰富的时空信息来实现强大的视觉跟踪。在[38]中，作者提出了一种具有注意力机制的新型跟踪框架，该机制通过选择相关过滤器的子集以提高鲁棒性和计算效率。滤波器的子集由深度注意力网络根据目标的动态属性进行自适应选择。该算法的贡献主要有：（1）引入注意力相关过滤器网络，该网络可以自适应跟踪动态目标。（2）利用注意力网络将注意力转移到最佳候选模块，并预测当前非活动模块的估计准确性。（3）扩大了相关滤波器的种类，涵盖目标漂移，模糊，遮挡，缩放变化和灵活的宽高比。（4）通过大量实验验证了视觉跟踪注意机制的鲁棒性和效率。

i) 基于串并联或级联结构的跟踪器

SPM-Tracker[39]的基本思想是在两个单独的匹配阶段解决两个需求。在粗匹配（CM）阶段，通过通用训练增强了鲁棒性，而在精细匹配（FM）阶段，通过在线学习，增强了网络辨别力。 FM 阶段的输入区域由 CM 阶段生成，因此这两个阶段串联连接。同时，这两

个阶段也被并行连接，因为匹配分数信息和目标边框位置信息被融合以生成最终结果。这种创新的串并联结构充分利用了两个阶段的优势，并具有出色的性能。在[40]中，作者指出，最近流行的 SiamRPN 目标跟踪器在存在表观近似的干扰目标和目标剧烈变化的情况下会退化。为了解决这些问题，作者提出了一个多阶段跟踪框架，即孪生级联 RPN（C-RPN），该框架由一系列来自孪生网络中不同层次的 RPN 组合而成。与以前的解决方案相比，C-RPN 具有几个优点：（1）每个 RPN 在上一阶段都使用 RPN 的输出进行训练。这样的过程会关注难分的负采样，从而使训练样本更加均衡。因此，RPN 在区分困难的背景（即类似的干扰因素）时将更具区分性。（2）提出了特征转移块（FTB）以充分利用多级特征，从而进一步使用高级语义和低级空间信息来改善 C-RPN 的辨别能力。（3）通过多步回归，C-RPN 在多个阶段逐步调整每个 RPN 中目标的位置和形状，从而使定位更加准确。

三、学位论文重要研究进展及成果描述

三.1 操纵模板像素以进行孪生视觉跟踪的模型自适应

尽管孪生网络在目标跟踪任务上取得了良好的性能，但是对于不包含在离线训练集中的目标，孪生网络所学习的相似性度量不一定是可靠的，从而导致泛化性较差。

我们通过直接操纵模板像素为孪生跟踪器提供一种新的模型自适应方法。我们首先回顾与基于模板匹配的跟踪器的跟踪过程，这与所提出的方法密切相关。我们将目标跟踪形式化为基于置信度的回归问题，该问题学习函数$s_\theta: \mathcal{Y} \times \mathcal{X} \to \mathcal{R}$，并根据给定的输出-输入对$(y, x)$预测标量置信度得分$s_\theta(y, x) \in \mathcal{R}$。最终估计值$f(x) = y^*$预测如下：

$$f(x) = arg \max_{y \in \mathcal{Y}} s_\theta(y, x), \tag{1}$$

其中x是输入图像。y通常表示目标中心对应的 2D 图像坐标。当前，有两种流行的基于模板匹配的范例：判别相关滤波器（DCF）方法和孪生跟踪方法。在基于 DCF 的方法中，在跟踪过程中训练一个循环相关滤波器w_θ以预测目标置信度得分：

$$s_\theta(y, x) = \big(w_\theta * \phi(x)\big)(y), \tag{2}$$

其中$\phi(x)$是从搜索图像x中提取的特征。与 DCF 相比，孪生跟踪器采用双流体系结构。一个流基于根据真实边界框从第一帧裁剪的模板图像z提取目标特征$\phi_\theta(z)$。另一流接收搜索图像x作为输入，并输出搜索特征$\phi_\theta(x)$。两个输出通过互相关以预测目标置信度得分：

$$s_\theta(y,x) = \big(\phi_\theta(z) * \phi_\theta(x)\big)(y). \tag{3}$$

基于 DCF 的跟踪器和孪生跟踪器均具有利用大规模跟踪数据集训练特征提取器$\phi(\cdot)$或嵌入网络$\phi_\theta(\cdot)$的优势，从而可以增强特征在目标跟踪任务上的表示能力。DCF 从目标图像块中学习滤波器w_θ，以将其与背景区分开。尽管使用了循环相关运算提高了跟踪效率，但其边界效应和复杂的优化却无法在计算速度和跟踪性能之间做出良好的权衡。孪生跟踪器在这方面做得更好，但在互相关中学习的相似性度量对于离线训练集中未包含的目标不一定是可靠的，从而导致泛化不佳。在这项工作中，我们旨在设计一种新的孪生跟踪方法，该方法能够充分利用当前视频的特定信息以进行模型调整，这是通过利用第一帧中的标注信息来实现的。注意公式（2）和公式（3）之间有一些相似之处，主要区别在于核的不同：DCF 的核是在线学习的w_θ，而孪生网络的核是$\phi_\theta(z)$。为了使孪生网络具有模型自适应能力，我们需要使用当前视频的第一帧标注信息来自适应地调整$\phi_\theta(z)$。有两种自适应调整$\phi_\theta(z)$的设计方法：更改$\phi_\theta(\cdot)$或更改z。然而，改变$\phi_\theta(\cdot)$可能会导致繁琐的元学习设置，从而无法确保离线训练的嵌入空间的生成能力。相比之下，我们的解决方案是以简单的方式通过更改z来执行孪生跟踪器的模型自适应，即仅使用第一帧的目标 ground truth，仅在几次梯度下降迭代中修改模板像素。与当前的孪生跟踪器模型自适应方法相比，该方法具有以下优点。首先，我们不修改孪生网络的参数，从而保留了离线训练的嵌入空间的表示能力。其次，与当前流行的模板更新方法，我们专注于在第一帧使用目标的 ground truth 进行初始自适应。最后，我们的模型自适应方法是即插即用的，因为它不会改变基本跟踪器的总体架构。接下来，我们将展示如何使用流行的对抗样本生成方法来执行用于模型自适应的模板像素操纵。

乍看之下，模型自适应任务与对抗样本生成任务之间没有直接的关联，因为这两个任务具有不同的目的。对抗样本指对输入数据进行非常轻微的修改，旨在对机器学习系统进行攻击，导致机器学习模型

做出错误的预测。而目标跟踪中模型自适应的目的是充分利用第一帧中的注释信息，以提高当前视频的跟踪性能。在下文中，我们将指出这两个任务之间存在一些相似之处，并且我们可以利用对抗样本生成方法来执行模型自适应任务。在介绍提出的方法之前，我们首先回顾流行的对抗样本生成方法。生成对抗图像的最简单方法之一是 FGSM，通过在干净样本的邻域中线性化损失函数，并通过闭合形式的方程找到精确的线性化函数的最大值：

$$I^{adv} = I + \epsilon\, sign\big(\nabla_I L(I, y_{true})\big), \tag{4}$$

其中I是输入图像，其像素值是$[0，\ 255]$范围内的整数。y_{true}是图像I的真实标签。$L(I,y)$是给定图像I和标签y的神经网络的成本函数。ϵ是要选择的超参数。扩展上述方法的一种直接方法是以较小的步长多次应用它，并在每个步骤之后裁剪中间结果的像素值，以确保它们位于原始图像的附近。这就是基本迭代方法（BIM）：

$$I_0^{adv} = I,\ I_{N+1}^{adv} = Clip_{I,\epsilon}\{I_N^{adv} + \alpha\, sign\Big(\nabla_I L\big(I_N^{adv}, y_{true}\big)\Big)\}, \tag{5}$$

其中$Clip_{I,\epsilon}\{\cdot\}$是对图像$I$执行像素值裁剪。前两种方法攻击属于无目标攻击，也就是所攻击方的目标仅仅是使得分类器给出错误预测，具体是哪种类别产生错误并不重要。BIM 可以轻松改进为有目标攻击，在这种情况下，攻击方想要将预测结果改变为某个指定的目标类别：

$$I_0^{adv} = I, I_{N+1}^{adv} = Clip_{I,\epsilon}\{I_N^{adv} - \alpha\, sign\Big(\nabla_I L\big(I_N^{adv}, y_{target}\big)\Big)\}. \tag{6}$$

上式表明，仅需进行几次梯度下降迭代操作就可以将输入图像的预测类别更改为y_{target}。注意到我们的目的是在第一帧中修改模板图像的像素，以便使预测更接近于真实边界框。因此，我们可以对有目标攻击方法进行简单的修改，用于对孪生网络的模型自适应：

$$z_0 = z,\ z_{N+1} = Clip_{z,\epsilon}\{z_N - \alpha\ sign\big(\nabla_z L(z_N, y_{bb})\big)\}, \tag{7}$$

其中z是第一帧中的模板图像，而y_{bb}是根据 ground truth 边界框生成的孪生跟踪网络的标签。接下来，我们将介绍带有自适应模块的跟踪器的整体跟踪过程。

提出的模型自适应方法以即插即用的方式与 SiamFC++跟踪器集成在一起。原始 SiamFC++网络的输入包括从第一帧裁剪的模板图像z_0和从第i帧裁剪的搜索图像x_i组成。但是，我们希望执行模板自适应，即利用输入对(z_0, x_0)进行N步像素更新后获得z'，跟踪器在(z', x_i)上表现良好。为此，我们首先使用真实边界框从第一帧裁剪初始模板图像$z_0 \in \mathcal{R}^{3\times128\times128}$和初始搜索图像$x_0 \in \mathcal{R}^{3\times289\times289}$。然后，将$z_0$和$x_0$

发送到 SiamFC++网络以获得第一帧的跟踪预测。SiamFC++中的跟踪损失计算如下：

$$L = L_{cls} + L_{quality} + L_{reg}, \tag{8}$$

其中L_{cls}是 focal loss。$L_{quality}$是用于质量评估的二进制交叉熵（BCE）损失。L_{reg}是边界框回归的 IoU 损失。相对于模板z_0的梯度用于迭代更新模板像素。

三.2 基于全局感知机制的长期目标跟踪系统

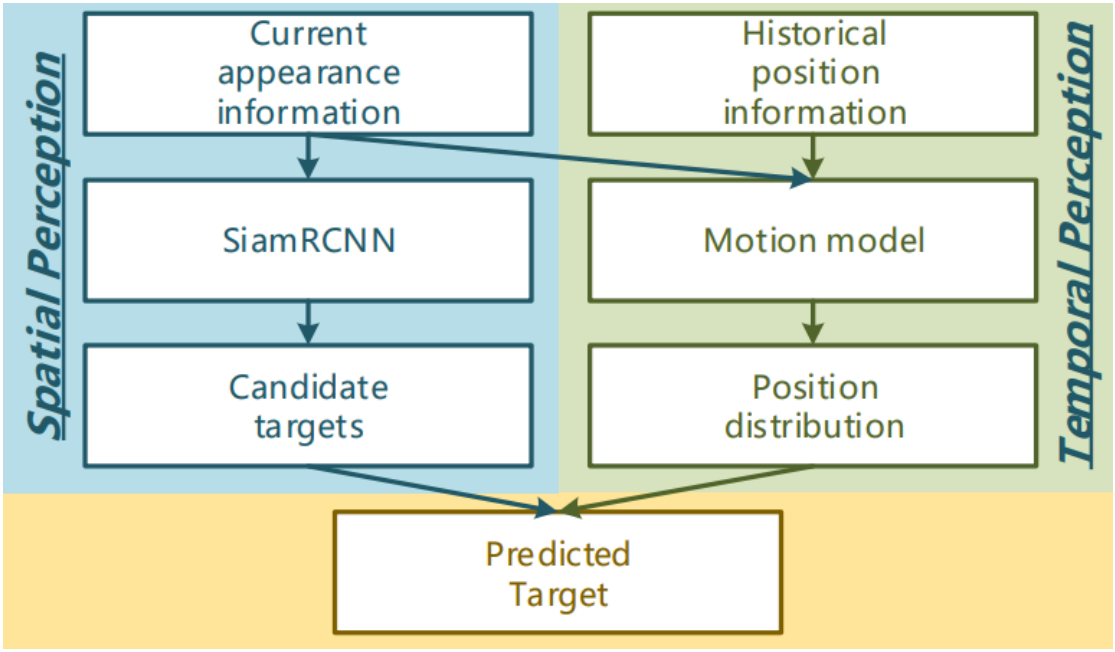

图 1：基于全局时空感知的跟踪框架。

受 Faster RCNN 的两阶段检测流程的启发，我们计划提出一种基于全局感知机制的孪生跟踪器。在跟踪过程中，所提出的跟踪器（即 SiamRCNN，图 2）始终能够感知整个图像上的目标。因此，即使跟踪器由于具有挑战性的目标表观变化而导致跟踪在当前帧出错，一旦其表观恢复正常，仍然可以及时检索到目标。尤其是在长期跟踪的情况下，当目标离开画面时，跟踪器无法在图像中找到目标，而当目标从任何位置重新进入屏幕时，我们的跟踪器可以继续工作。除了上述全局空间感知机制外，我们还计划提出一个运动模型来减轻相似目标的干扰。众所周知，基于孪生网络的跟踪框架常常会受到相似目标的困扰。与简单设计的手工策略不同，我们计划提出的基于 CNN 的运动模型经过端到端训练，可以使用历史轨迹信息和当前目标表观信息来预测目标在当前帧中的位置分布。具体来说，目标的运动模式是在

训练阶段从大规模轨迹数据集中自动学习的，而不是手工设计的特征或规则。在测试阶段，我们可以使用任意数量的历史帧的位置信息，而不仅仅是先前的帧进行预测。我们所提出的跟踪框架的的另一个优势是，它使用 RoI Align 操作来使跟踪器能够利用更深的网络进行特征学习，这对于基于局部机制的 SiamFC 和 SiamPRN 跟踪器来说是不可接受的，因为其网络结构中的填充将破坏严格的平移不变性。在我们的跟踪网络中，将跟踪器中的整个模板图像和搜索图像发送到相同的骨干网络以提取特征，然后根据初始帧的中的目标位置标注使用 RoI Align 操作从模板特征中获取目标特征。通过逐通道互相关操作使目标特征和搜索图像特征相互结合以获得融合特征图，该融合特征图被发送到后续的跟踪模块。由于此解决方案对平移不变性没有任何限制，因此我们可以使用功能强大的 ResNet 作为骨干网络来提高网络的学习能力。综上所述，该算法具有如下优势：（1）基于全局感知机制，我们提出一个两阶段跟踪框架，使用比常见孪生跟踪器更深的网络来减少累积误差并提高鲁棒性。 （2）提出一种基于端到端学习的卷积神经网络的轨迹预测模块，利用历史轨迹信息和当前帧的表观

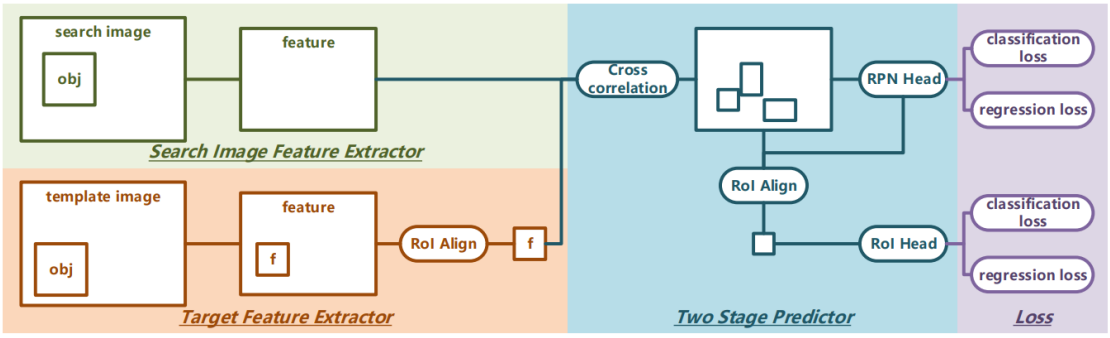

图 2：SiamRCNN 结构图。

信息来预测目标位置分布。

具体而言，计划将目标跟踪任务分解为以下步骤：首先提取候选目标（包括真实目标和位于背景的近似目标），然后使用运动模型排除近似目标。通过这种方式，我们可以设计更精确的跟踪器和更鲁棒的运动模型来获得更好的性能，尤其是在长期跟踪的场景中。为了解决由局部搜索引起的累积误差，我们计划提出全局时空感知跟踪系统，其内容包括：（1）使用整幅图像而不是小的图像块作为跟踪器的输入，从而为跟踪器提供全局空间信息。（2）为了更好地感知全局空间信息，我们计划提出 SiamRCNN 跟踪器，该跟踪器能够检测与真实目标表

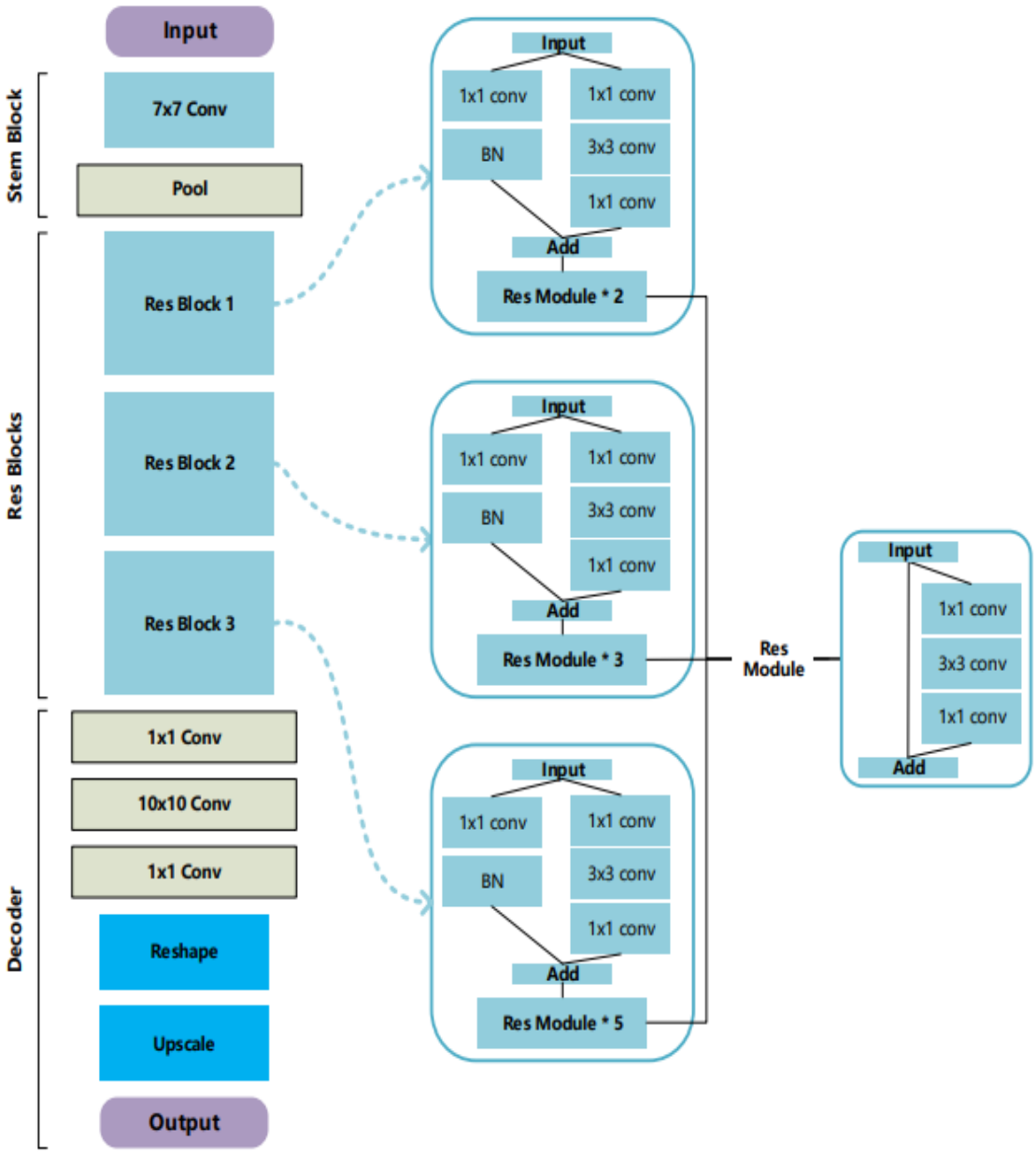

图 3：InstMask 网络结构图。

观近似的候选目标。（3）为了感知时间信息，我们计划提出一种运动模型，该模型能够通过预测位置分布来排除近似目标的干扰以获得最终的跟踪结果。大多数流行的跟踪器采用一阶段的框架，该框架基于通过互相关获得的特征执行分类和边界框回归。在物体检测领域中，两阶段检测器的性能通常优于一级检测器。受此启发，我们将跟踪器设计为两级网络。 RPN 阶段快速过滤掉大多数背景样本，并且 RoI 模块采用固定的前景与背景比率，以维持前景与背景之间的样本平衡。

此外，即使对于形状极端的目标，也可以通过两阶段回归来实现精确的定位。

SiamRCNN 跟踪器由四个模块组成：（1）特征提取模块，（2）特征融合模块，（3）RPN head 模块和（4）RoI head 模块。特征提取模块有两个输入：模板图像 z 和搜索图像 x。根据孪生体系结构的设计，两个输入共享相同的网络参数以提取特征。特征提取模块的网络结构是 ResNet50 的变体，它在 1000 类 ImageNet 分类数据库上进行了预训练。从 ResNet50 的最后一个卷积层中提取输入图像的特征。将获得的通道尺寸为 1024 和步长为 16 的模板特征$\varphi(z)$和搜索特征$\varphi(x)$发送到后续特征融合模块。

在特征融合模块中，根据目标的标注信息，通过 RoI Align 操作从模板特征中获取目标特征。搜索特征和目标特征通过逐通道互相关进行融合。在互相关层的顶部添加两个通道尺寸为 1024 的 1×1 卷积以获得融合特征。

RPN head 包括两个并联的 1×1 卷积层——通道尺寸为 2k 的分类层和通道尺寸为 4k 的回归层，其中 k 是每个位置的最大可能候选框数量。RPN head 将融合特征作为输入，并同时回归每个锚点的区域边框和分类得分。

在 RoI head 中，通过在融合特征上执行 RoI Align，为来自 RPN 的每个候选区域提取深度特征，从而为每个 RoI 生成一个小特征图，其通道尺寸为 2048，空间分辨率为 7×7。 由 RoI Align 得到的候选区域特征被送入到全局平均池化层，然后是两个并联的输出层：一个层在两个类别（前景或背景）上生成 softmax 概率估计值，另一个层为前景类输出四个实数值，用于编码 RoI 的精确边界框位置。

在使用我们的 SiamRCNN 检测到与给定的第一帧模板目标在视觉上相似的目标区域之后，我们使用运动模型消除近似目标并获得最终的跟踪结果。通过使用历史轨迹信息和当前帧的表观信息学习目标位置分布，运动模型以端到端的方式工作。通过运动模型预测的位置分布，反映了在每个空间位置出现目标的可能性。通过巧妙的设计，我们能够使用诸如 HRNet [45]之类的姿态估计网络来构建运动模型。可以将姿态估计网络用于设计运动模型的主要原因是，姿态估计网络的输出是关节点的位置分布，而运动模型的输出是目标的位置分布。这意味着这两个任务之间有很多共同点。为了使用 HRNet 进行轨迹估计，我们需要更改 HRNet 的输入和输出，并保持网络结构不变。在跟踪第 i 帧期间，我们利用前 K 帧的位置信息作为网络的输入。具体

而言，对于历史帧，生成以该帧中的目标位置为中心的标准差为 3 个像素的 2D 高斯生成热图。生成的 K 个热图根据时间顺序在通道上进行堆叠，以获得通道尺寸为 K 的轨迹张量。我们的运动模型不仅利用历史轨迹信息进行预测，而且还考虑了当前帧的表观信息。为此，将当前帧的 RGB 图像和轨迹张量进行级联，以获得通道尺寸为（3＋K）的张量。该张量发送到网络，网络的输出是一个热图，反映了目标在当前帧中的位置分布。网络的 ground truth 是以当前帧中的目标位置为中心的标准差为 3 个像素的 2D 高斯热图。我们的运动模型的网络结构与 HRNet 相同，其简要描述如下：HRNet 的第一阶段是高分辨率子网。然后，继续在网络中添加不同分辨率的子网以形成更多阶段。多分辨率子网之间并行连接。通过在并行多分辨率子网中多次信息交换以进行重复的多尺度融合。

三. 3 孪生跟踪器的端到端时间聚合

在孪生网络跟踪器中，由于视频的运动模糊，遮挡等原因可能导致物体的表观变差，学习到的物体特征可能没有那么大的判别力。以前的工作尝试过很多方法来改进特征表示。但是，大多数算法都基于仅利用当前帧的特征执行跟踪，这限制了孪生跟踪器的潜力。实际上，视频具有关于目标的丰富信息，并且这种时间信息是视频理解和跟踪的重要基础。在本文中，我们提出了一种新的孪生跟踪器 SiamTFA，通过聚合相邻帧的时间信息来改善每帧的特征。这种时间融合策略使暹罗跟踪器能够处理较差的目标表观，例如运动模糊，遮挡等。具体而言，SiamTFA 将图像对作为输入，包括模板图像和搜索图像。模板图像是根据真实边界框从初始帧裁剪的图像块。搜索图像是视频的某个后续帧。 两个输入共享相同的特征提取器和参数。受两阶段检测框架的启发，我们的孪生跟踪器也是一种两阶段方法。第一阶段旨在生成在视觉上类似于给定模板目标的候选区域。在这个阶段，我们引入一个时间聚合模块来增强时间信息。第二阶段旨在从候选框中确定目标。在这一阶段，我们插入一个对抗性 dropout 模块以学习更鲁棒的特征。

具体来说，候选生成阶段包括 3 个组件：（1）特征提取器，（2）时间聚合模块和（3）特征调制模块。特征提取器分别为搜索图像和模板图像生成搜索特征和模板特征。时间聚合模块被集成到特征提取

器中以利用时间信息。特征调制模块融合搜索特征和模板特征以识别候选目标

为了处理目标的比例变化，我们使用 Res50-FPN 作为特征提取器。特征金字塔网络（FPN）利用深层卷积网络固有的多尺度金字塔层次结构来构建特征金字塔。我们的孪生 FPN 将模板图像和搜索图像作为输入。对于搜索图像，FPN 以完全卷积的方式在多个级别上输出不同比例大小的特征图。我们将搜索图像的输出表示为$F_x = \{f_x^i\}_{i=1:5}$，它们相对于输入搜索图像具有{4、8、16、32、64}像素的步幅。对于模板图像，我们将 FPN 输出的最后阶段用作模板特征，其空间大小为 7×7。

孪生网络跟踪器通常使用单帧图像进行跟踪结果的预测。这限制了孪生跟踪器的能力。一方面，单帧跟踪会产生不稳定的结果，并且在表观不佳时会失败。另一方面，在时间上相邻的帧可以提供有关目标的更多信息。因此，我们旨在通过汇总相邻帧的特征来改善每帧特征。具体来说，我们将一个时间聚合模块插入特征提取器的最后一个阶段。为了建模时间信息，在网络训练过程中，一个 batch 中的图像是同一视频中的几个相邻帧，并按时间排序，因此我们可以将 batch 维度视为时间维度。假设特征提取器最后阶段的特征图为$f \in \mathcal{R}^{T\times C\times H\times W}$。对于每个时刻$t \le T$，我们首先将特征$f^t \in \mathcal{R}^{C\times H\times W}$沿通道维度分为 3 个部分：$f_{1:K}^t \in \mathcal{R}^{K\times H\times W}$，$f_{(K+1):2K}^t \in \mathcal{R}^{K\times H\times W}$和$f_{(2K+1):C}^t \in \mathcal{R}^{(C-2K)\times H\times W}$。然后我们沿着时间维度移动通道：

$$f_{agg}^t = \mathcal{C}\left(f_{1:K}^{t-1}, f_{(K+1):2K}^{t+1}, f_{(2K+1):C}^t\right), \tag{8}$$

其中$\mathcal{C}(\cdot)$是串接操作。我们仅在残差层执行移位操作以保留孪生跟踪器的空间特征学习能力。请注意，聚合特征f_{agg}^t与f^t具有相同的形状，因此我们可以将此模块直接插入到主干中，而无需更改网络的其他部分。而且，此操作仅需要进行数据移动，因此不消耗额外的计算量，并且可以进行端到端的培训。

在获得模板特征f_z和搜索特征金字塔$F_x = \{f_x^i\}_{i=1:5}$之后，对它们进行调制以生成特定于目标的特征。具体而言，调制向量f_{avg}是将f_z进行全局平均池化得到的，用于表示特定目标的表观信息。调制特征金字塔$F_{mod} = \{f_{mod}^i\}_{i=1:5}$生成方式如下：

$$f_{mod}^i = \mathcal{M}\left(f_{avg}, f_x^i\right), \tag{9}$$

其中$\mathcal{M}(\cdot)$是逐通道互相关。每个调制特征图都送入两个的全连接层中，即通道尺寸为$4k$的边框回归层和通道尺寸为$2k$的边框分类层，其中k是每个位置的最大可能候选数。

三.4 基于实例引导的相关滤波器跟踪算法研究

传统的相关滤波跟踪器往往依赖底层像素信息进行学习，缺乏高级语义信息的引导。为了利用目标实例级别的语义信息约束滤波器的学习，我们提出了实例引导的相关过滤器（IGCF），采用一个深层网络（即 InstMask）旨在生成目标的准确分割模板。InstMask 经过离线的端到端的训练，可以从 COCO2017[46]中学习语义信息。实例分割模板可以通过抑制背景杂波的干扰来显式限制相关滤波器的学习过程。与常见的目标分割任务不同，我们设计了一种新的网络结构和训练方法，使 InstMask 能够识别位于搜索图像块中心的任意类别的显著目标。这个轻量级的网络在单个 CPU 核心平台上以 5 FPS 的速度

Layers	Output Size	Support
Data	$160 \times 160 \times 3$	-
Stem Block	$80 \times 80 \times 64$	7×7 conv
	$40 \times 40 \times 64$	2×2 maxpool
Res Block (1)	$40 \times 40 \times 256$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 1 \times 1 \text{ conv} \end{bmatrix} \times 3$
Res Block (2)	$20 \times 20 \times 512$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 1 \times 1 \text{ conv} \end{bmatrix} \times 4$
Res Block (3)	$10 \times 10 \times 1024$	$\begin{bmatrix} 1 \times 1 \text{ conv} \\ 3 \times 3 \text{ conv} \\ 1 \times 1 \text{ conv} \end{bmatrix} \times 6$
Decoder	$10 \times 10 \times 128$	1×1 conv
	$1 \times 1 \times 512$	10×10 conv
	$1 \times 1 \times 3136$	1×1 conv
	56×56	reshape
	160×160	upscale

表 1：InstMask 网络结构表。

运行，在跟踪精度和速度之间实现了平衡。此外，我们注意到，为满

足跟踪自适应性要求而在线更新的相关滤波器与满足鲁棒性要求的静态 InstMask 模块是独立且互补的。因此，两个组件的输出可以集成在一起，以进一步提高跟踪性能。具体来说，基于实例级别的分割，我们进一步提出了一种自校正机制来减轻相关滤波跟踪器的漂移问题。分割模板的几何中心用于校正相关滤波器的预测偏差。配备了 InstMask 模块的 IGCF 不仅可以精确地跟踪目标，还可以用于视频目标分割，这证明了我们算法的广泛应用。

出于效率和性能方面的考虑，理想的分割模块应具有三个关键特征：（1）与跟踪过程可以无缝集成；（2）足够简单以适应跟踪的速度要求；（3）所提出的分割网络应尽可能精确分割物体的轮廓。以前大多数用于生成空间约束的方法都依赖于手工设计的规则或低级图像特征（例如颜色直方图）。在本文中，与以前的生成分割模板的方法不同，我们不依赖于边缘，超像素或任何其他形式的低级分割特征。相反，我们模型的核心是卷积神经网络。通过利用在 ImageNet[47]上训练的强大卷积特征表示并在 COCO 数据集上进行微

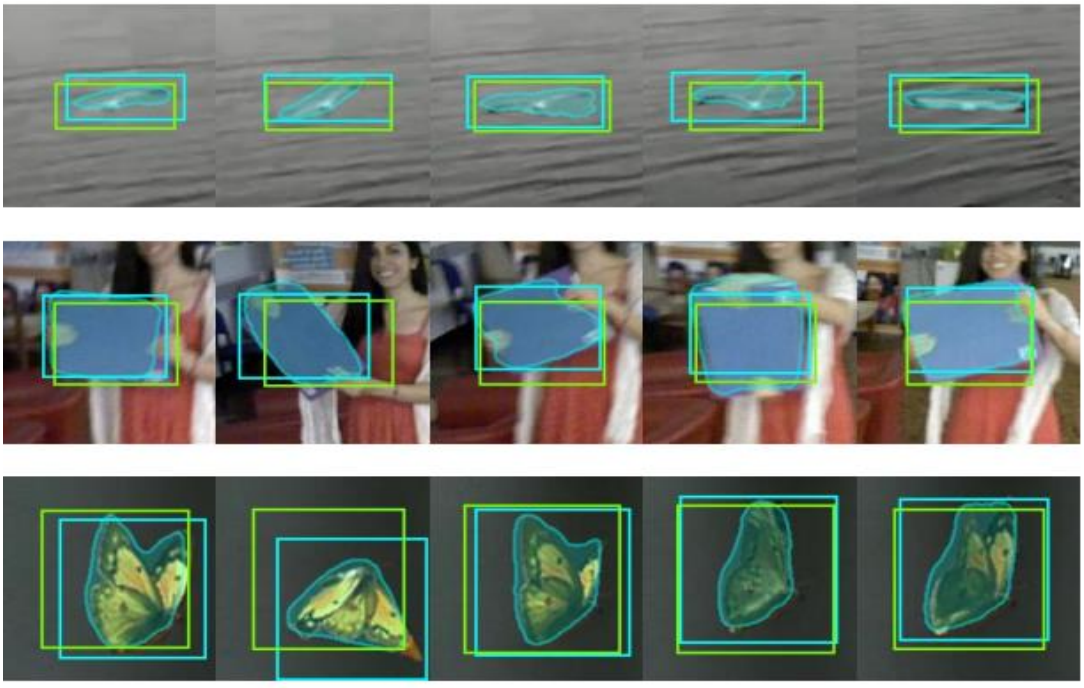

图 4：使用 IGSC 模块之前（绿色）和之后（蓝色）的跟踪结果。

调，我们能够为目标生成语义分割模板，以约束相关跟踪器的学习。值得注意的是，将现有的实例分割方法[48, 49, 50]进行集成

来为相关过滤器学习提供空间约束是次优的。通用跟踪算法通常具有复杂的网络结构，因此很难在跟踪精度和速度之间取得平衡。相比之下，本文提出的轻量化分割网络是专为跟踪而设计的，在 CPU 上以 5 FPS 运行。在第 i 帧中进行跟踪时，需要获取第 i 帧中目标的分割模板以限制相关滤波器的学习。假设第 i 帧中的目标位置在第 i-1 帧中的目标位置附近，那么对 InstMask 的输入始终是以前一帧的目标位置为中心的小图像块。该设计具有以下优点：（1）由于输入数据的强一致性，网络可以使用较少的参数获得准确的分割结果。（2）由于网络参数较少且图像补丁较小，因此推理所需时间较短。（3）在目标的先前位置附近进行搜索可避免位于背景的相似目标的干扰。在训练期间，训练集中的每个样本包含（1）RGB 图像块，其中包含一个目标，该目标靠近输入图像块的中心，（2）对应于输入图像的二进制分割模板。InstMask 的网络结构网络体系结构如图 3 所示。InstMask 的骨干网络是基于 ResNet50 构建的，该网络由一个 stem 块和 3 个残差块组成。InstMask 的输入是 160×160×3 的图像块，被送入 stem 块，生成 40×40×64 特征图。然后将特征图发送到 3 个残差块，分别生成大小为 40×40×256、20×20×512、10×10×1024 的特征图。随后，将获得的特征图发送到三个卷积层以生成长度为 3136 的向量，并重新调整为 56×56，以获得最终的分割模板。由于该分割网络是一个轻量级的网络，因此可以在具有 Intel E5-2620 CPU 内核的平台上以 5 FPS 的速度运行，在跟踪精度和速度之间实现最佳平衡，从而可以将跟踪器部署在包括自动驾驶，机器人技术和增强现实等实际应用中。在网络训练期间，尺寸为 56×56 的分割模板被上采样到原始图像尺寸 160×160。令L表示尺寸为$w \times h$的像素级 ground truth 分割模板，则损失计算如下：

$$\mathrm{L} = \frac{1}{wh}\sum_{xy} log\left(1 + e^{-l_{x,y}P_{x,y}}\right) \tag{10}$$

COCO2017[46]实例分割数据集用于训练 InstMask。与仅预测出现在训练集中的类别的许多分割网络相反，提出的 InstMask 在训练期间忽略类别信息，以训练类别无关的分割网络。实际上，InstMask 能够检测到位于搜索区域的显著目标。尽管 InstMask 使用只有 80 个类别的 COCO 数据集进行训练，但是网络能够对训练集中未出现的类别进行分割。我们使用随机梯度下降训练模型，批处理大小为 32 个样本，动量为 0.9，权重衰减为 0.00005。总共有 50

个 epochs。在对数空间中，学习率从10^{-2}降低到10^{-4}。在训练过程中，输入图像大小设置为 160×160，目标位于图像中心，大小为 112×112。为了增强网络的通用性，我们执行数据增强。具体来说，我们考虑平移（±16 像素），缩放变形（2±1/4）以及水平翻转。

InstMask 是离线训练的。在跟踪期间，InstMask 仅执行前向传播过程，而没有梯度的反向传播。这不仅可以防止漂移，还可以满足跟踪的速度要求。尽管我们的模型与 CSRDCF 具有相似性，但其实现原理却大不相同。在 CSRDCF 中，使用目标和背景的颜色直方图生成分割模板。这种启发式方法具有几个缺点：（1）由于仅使用低级像素信息，因此难以准确地分割目标。（2）为了适应目标的明显变化，在跟踪过程中会不断更新直方图模型，这很容易导致跟踪器漂移。相反，InstMask 不会在线更新参数。这提高了计算效率，同时避免了不希望的漂移。此外，由于使用语义级别而不是像素级别的信息进行分割，因此跟踪结果更加准确。

实例引导的自校正组件：InstMask 有利于学习更好的滤波器，并有助于纠正不良的跟踪结果。在 DCF 模块中，滤波器h和特征f之间的相关值最大的位置被认为是当前帧中的目标位置。但是，由于在线更新，相关滤波器很容易漂移。 另一方面，InstMask 的结果是从 COCO 数据集中学习得到的。但是，我们不能仅依靠目标的语义分割模板来产生最终的跟踪结果，这是由于 InstMask 不在线更新。为了利用两个结果的优点并克服它们的缺点，我们提出了实例指导的自校正组件 IGSC。IGSC 将这两个结果结合起来，以获得更好的跟踪：

$$m_{x,y} = \begin{cases} 1, \text{if } P_{x,y} > b \\ 0, \text{if } P_{x,y} \le b \end{cases} \tag{11}$$

其中P是模板热图，b是 0 到 1 之间的阈值，m是二值分割模板。则目标位置可以通过下式获得：

$$c_m = Centroid(m) \tag{12}$$

其中 $Centroid(\cdot)$可以计算区域的几何中心。将 p 表示为校正后的目标位置。自校正过程可以描述为：

$$p = \begin{cases} c_m, \text{if } |c_m - c_{dcf}|_2^2 < \beta \\ c_{dcf} + \alpha \cdot (c_m - c_{dcf}), \text{otherwise} \end{cases} \tag{13}$$

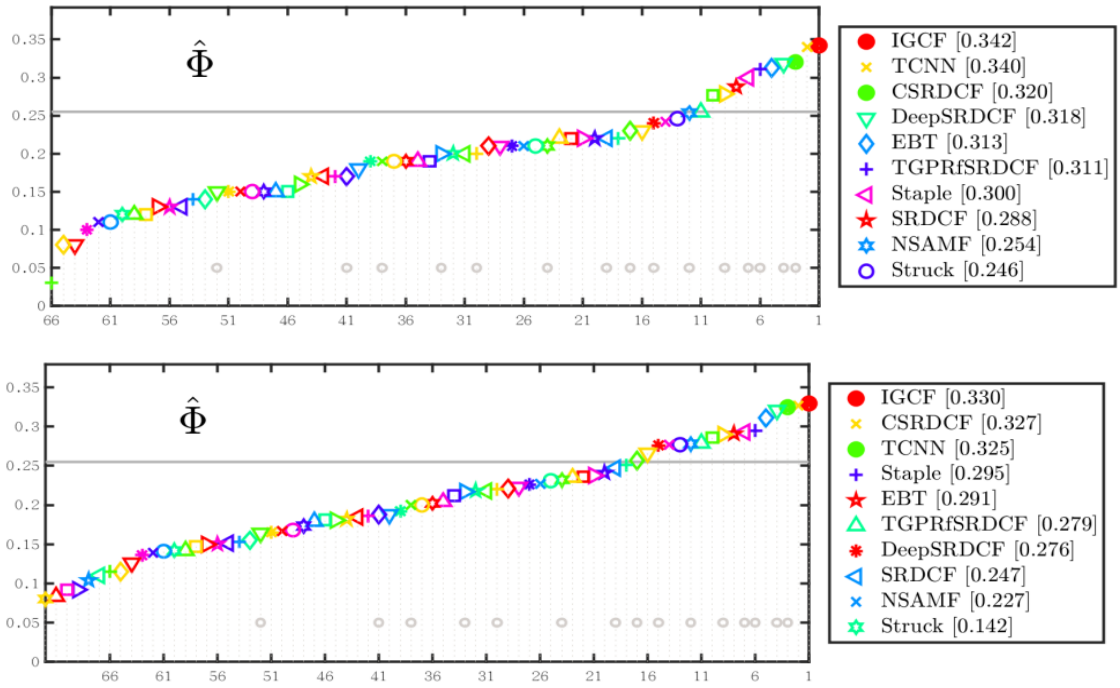

图 5：IGCF 在 VOT2015（上）和 VOT2016（下）的跟踪结果（EAO）。

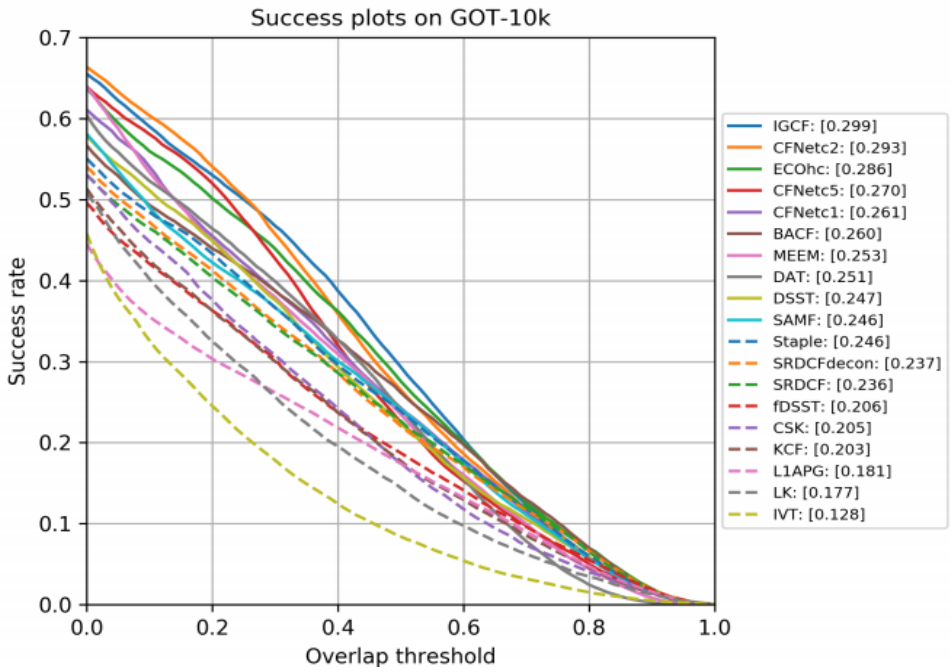

图 6：IGCF 在 GOT10k 上的跟踪结果。

其中c_{dcf}是 DCF 预测的位置，α是控制自校正强度的超参数，β是c_m与c_{dcf}之间距离的阈值。由于分割结果的鲁棒性，当 DCF 结果显示

不稳定的漂移时，跟踪器可以自行校正。所提出的实例指导的自校正组件的有效性如图 4 所示。

我们在具有挑战性的三个目标跟踪数据集上：（VOT2015，VOT2016 和 GOT-10k），对提出的跟踪器 IGCF 进行了全面评估。请注意，用于训练 InstMask 组件的视频与评估数据集之间没有重叠。另外，对视频对象分割数据集 DAVIS 进行了定性实验。

在 VOT2015 的实验结果：VOT2015 跟踪数据集包含 60 个具有挑战性的视频。它是通过先进的序列选择方法从 300 多个序列中构建的，该方法选择了难以跟踪的目标并最大化了视觉属性多样性成本函数。在 VOT2015 中，使用三个指标来评估跟踪器的性能：（1）准确性，（2）鲁棒性和（3）EAO。准确性衡量预测的边界框与 ground truth 的重叠程度。鲁棒性衡量跟踪器在跟踪过程中失去目标的次数。EAO 结合了准确性和鲁棒性来评估跟踪器的整体性能。我们将算法与以下跟踪器进行了比较：CSRDCF，SRDCF，TGPRfSRDCF，DeepSRDCF，TCNN，Staple，EBT，NSAMF 和 Struck。这些跟踪器可以分为三类：Struck 和 EBT 是传统的跟踪器。SRDCF，CSRDCF，TGPRfSRDCF，Staple 和 SAMF 是基于 DCF 的跟踪器。DeepSRDCF 和 TCNN 是基于 CNN 的跟踪器。跟踪器的 EAO 得分如图 5 所示。与其他列出的方法相比，我们的方法可实现 0.342 的 EAO。 Struck 使用内核化的结构化输出支持向量机（SVM），可以在线学习以提供自适应跟踪。相反，IGCF 基于强大的相关过滤器。因此，就 EAO 而言，IGCF 大大超过了 Struck，其相对增幅为 36.8%，这表明 DCF 框架的出色表现以及 IGCF 的有效性。CSRDCF 使用目标和背景的颜色直方图生成目标模板，以限制相关滤波器的学习，而 IGCF 使用神经网络生成模板。与 CSRDCF 相比，EAO 分数的相对提升为 6.9%，这表明，与从低级图像特征生成的模板相比，由深层特征生成的语义掩码更有利于过滤器学习。DeepSRDCF 在 SRDCF 框架中结合了来自预训练网络的深度特征。但是，这些深度特征并不具有显示的语义信息。 相反，InstMask 是针对 COCO 分割数据集进行训练的，该数据集是为实例分割而设计的。在 VOT2015 上，就 EAO 而言，IGCF 比 DeepSRCDF 高出约 7.5%，这反映了 InstMask 和 IGSC 模块的重要性。

在 VOT2016 的实验结果：VOT2016 数据集包含来自 VOT2015 的 60 个具有改进注释的序列。VOT2016 的评估指标与 VOT2015 相同，即准确性，鲁棒性和 EAO。我们将我们的算法与以下跟踪器进行了

比较：CSRDCF，SRDCF，TGPRfSRDCF，DeepSRDCF，TCNN，Staple，EBT，NSAMF 和 Struck。图 5 显示了在 VOT2016 上 EAO 的性能。在列出的方法中，我们的方法以 0.330 的 EAO 得分达到最佳结果。

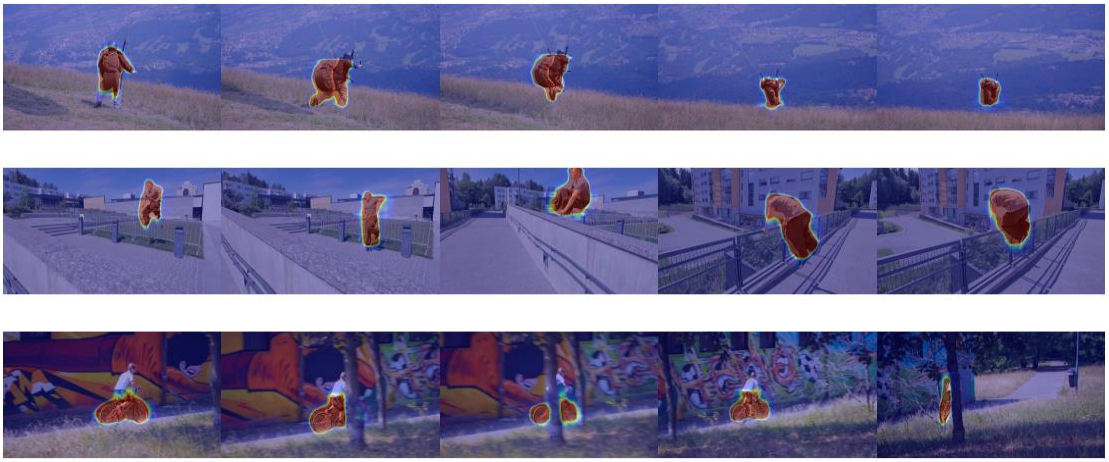

图 7：IGCF 在 DAVIS2016VAL 上的定性结果。

对 GOT-10k 的评估：GOT-10k 是一个用于通用对象跟踪的大型数据库。该数据集包含超过 10000 个真实运动对象的视频片段和超过 150 万个手动标记的边界框，对应于现实世界中的 563 类运动对象和 87 种运动模式。据我们所知，GOT-10k 是迄今为止最丰富的运动轨迹数据集。我们在 GOT-10k 测试子集上评估跟踪器，该子集具有 180 个视频段，包含 84 个物体类别和 32 个运动模式。GOT-10k 的评估指标是平均重叠（AO），它表示 ground truth 和预测的边界框之间的重叠平均值。我们将我们的算法 IGCF 与以下跟踪器进行了比较：CFNetc2，ECOhc，CFNetc5，CFNetc1，BACF，MEEM，DAT，DSST，SAMF，Staple，SRDCFdecon，SRDCF，fDSST，CSK，KCF，L1APG，LK 和 IVT。跟踪结果在图 6 中。我们的方法的 AO 得分为 0.299，这比所有其他列出的竞争性跟踪算法都要好。

对 DAVIS2016 的评估：由于具有 InstMask 模块，所提出的跟踪器不仅可以很好地进行目标跟踪，还可以用于视频分割。我们展示了在 DAVIS2016 VAL 数据集中的定性结果。DAVIS2016 是一个视频目标分割数据集，它由五十个高质量的视频序列组成。可视化结果如图 7 所示。

四、学位论文工作进度安排

2020 年 12 月—2021 年 01 月：完善基于实例引导的相关滤波器跟踪算法研究工作，发表至相关期刊或会议论文。

2021 年 02 月—2021 年 04 月：整理相关科研成果，根据总结的大纲，撰写博士学位论文。

2021 年 05 月—2021 年 06 月：根据博士学位论文，准备对应的答辩 PPT，进行毕业答辩。

五、课程主要完成情况

学年学期	课程名称	学时	学分	成绩	学位课
2016—2017学年(秋)第一学期	模式识别与机器学习	60	3.0	87	是
	人工智能理论与实践	60	3.0	96	是
	图像处理与分析	60	3.0	90	是
	矩阵分析与应用	40	2.0	90	是
	随机过程	40	2.0	81	否
	中国马克思主义与当代	36	1.0	82	是
	人文系列讲座	20	1.0	通过	是
	中国特色社会主义理论与实践研究	36	1.0	82	是
	论语研读	40	1.0	83	否
	博士学位英语（免修）	72	2.0	75	是
	硕士学位英语（免修）	72	3.0	74	是
	英文科技论文写作	32	1.0	88	否
	体育类公共选修课	26	0.5	通过	否
2016—2017学年(春)第二学期	世界艺术与建筑史话	30	1.0	84	否
	中国民居建筑艺术赏析	30	1.0	83	否
	计算机视觉	40	2.0	94	是
	生物特征识别	32	2.0	83	是
	视频处理与分析	40	2.0	87	是
	自然辩证法概论	36	1.0	82	是
2016—2017学年(夏)第三学期	网路安全导论	20	1.0	通过	否
	统计机器学习理论	20	1.0	93	否
2017—2018学年(秋)第一学期	积极心理学	20	1.0	87	否
2018—2019学年(秋)第一学期	最优化算法理论与应用	42	2.0	70	是
	人工智能前沿讲座	24	1.0	80	是
	以下空白				
总学分	38.5		学位课学分		29.0

六、已取得的阶段性成果

[1] Zhenbang Li, Qiang Wang, Jin Gao, Bing Li, Weiming Hu. Globally Spatial-Temporal Perception: A Long-Term Tracking System. IEEE International Conference on Image Processing, 2020, Accepted. (EI, CCF=C)

[2] Zhenbang Li, Qiang Wang, Jin Gao, Bing Li, Weiming Hu. End-To-End Temporal Feature Aggregation For Siamese Trackers. IEEE International Conference on Image Processing, 2020, Accepted. (EI, CCF=C)

[3] Zhenbang Li , Bing Li , Jin Gao , Liang Li, Weiming Hu. Manipulating Template Pixels for Model Adaptation of Siamese Visual Tracking. IEEE Signal Processing Letters, 2020, Accepted. (SCI, JCR Q2)

七、学位论文开题存在的问题及回复

问题：开题报告中提及的“实现基于大数据驱动的运动模型”一句中，“大数据”一词表述不当。

回复：正确理解“大数据”一词的含义。大数据并非简单地表示大规模数据集，而是指一种规模大到在获取、存储、管理、分析方面大大超出了传统数据库软件工具能力范围的数据集合，具有海量的数据规模、快速的数据流转、多样的数据类型和价值密度低四大特征。因此，将“实现基于大数据驱动的运动模型”表述修改为“实现基于CNN的运动模型，从大规模轨迹数据集中学习物体的运动模式”。

八、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，以及自己的研究工作与它们的联系。

第三章介绍我的第一个研究工作：操纵模板像素以进行孪生视觉跟踪的模型自适应。

第四章介绍我的第二个研究工作：基于全局感知机制的长期目标跟踪系统。

第五章介绍我的第三个研究工作：孪生跟踪器的端到端时间聚合。

第六章介绍我的第四个研究工作：基于实例引导的相关滤波器跟踪算法研究。

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。

附：主要参考文献

[1] Smail H, David H, Larry S D. W4: Real-Time surveillance of People and their Activities[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(8).

[2] Collins R T, Lipton A J, Kanade T, et al. A system for video surveillance and monitoring[J]. VSAM final report, 2000: 1-68.

[3] Coifman B, Beymer D, McLauchlan P, et al. A real-time computer vision system for vehicle tracking and traffic surveillance[J]. Transportation Research Part C: Emerging Technologies, 1998, 6(4): 271-288.

[4] Masoud O, Papanikolopoulos N P. A novel method for tracking and counting pedestrians in real-time using a single camera[J]. IEEE transactions on vehicular technology, 2001, 50(5): 1267-1278.

[5] Ma C, Huang J B, Yang X, et al. Hierarchical convolutional features for visual tracking[C]//Proceedings of the IEEE international conference on computer vision. 2015: 3074-3082.

[6] 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.

[7] Song Y, Ma C, Wu X, et al. Vital: Visual tracking via adversarial learning[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8990-8999.

[8] Wang X, Li C, Luo B, et al. SINT++: robust visual tracking via adversarial positive instance generation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4864-4873.

[9] Gao J, Zhang T, Xu C. Graph convolutional tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 4649-4659.

[10] Tu Z, Zhou A, Jiang B, et al. Visual Object Tracking via Graph

Convolutional Representation[C]//2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2019: 234-239.
[11] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[C]//Advances in neural information processing systems. 2014: 3104-3112.
[12] Donahue J, Anne Hendricks L, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 2625-2634.
[13] Du Y, Wang W, Wang L. Hierarchical recurrent neural network for skeleton based action recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1110-1118.
[14] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 129-136.
[15] Pinheiro P H O, Collobert R. Recurrent convolutional neural networks for scene labeling[C]//31st International Conference on Machine Learning (ICML). 2014 (CONF).
[16] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//2013 IEEE international conference on acoustics, speech and signal processing. IEEE, 2013: 6645-6649.
[17] Graves A, Schmidhuber J. Offline handwriting recognition with multidimensional recurrent neural networks[C]//Advances in neural information processing systems. 2009: 545-552.
[18] Shuai B, Zuo Z, Wang G. Quaddirectional 2d-recurrent neural networks for image labeling[J]. IEEE Signal Processing Letters, 2015, 22(11): 1990-1994.
[19] Li Y, Zhu J, Hoi S C H. Reliable patch trackers: Robust visual tracking by exploiting reliable patches[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 353-361.
[20] Stollenga M F, Byeon W, Liwicki M, et al. Parallel multi-dimensional LSTM, with application to fast biomedical volumetric image segmentation[C]//Advances in neural information processing systems. 2015: 2998-3006.
[21] Byeon W, Breuel T M, Raue F, et al. Scene labeling with lstm recurrent neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3547-3555.
[22] Tai K S, Socher R, Manning C D. Improved semantic representations

from tree-structured long short-term memory networks[J]. arXiv preprint arXiv:1503.00075, 2015.
[23] Zhu X, Sobihani P, Guo H. Long short-term memory over recursive structures[C]//International Conference on Machine Learning. 2015: 1604-1612.
[24] Cui Z, Xiao S, Feng J, et al. Recurrently target-attending tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1449-1458.
[25] Gordon D, Farhadi A, Fox D. Re $^ 3$: Re al-Time Recurrent Regression Networks for Visual Tracking of Generic Objects[J]. IEEE Robotics and Automation Letters, 2018, 3(2): 788-795.
[26] 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.
[27] 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.
[28] Yun S, Choi J, Yoo Y, et al. Action-decision networks for visual tracking with deep reinforcement learning[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2711-2720.
[29] Supancic III J, Ramanan D. Tracking as online decision-making: Learning a policy from streaming videos with reinforcement learning[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 322-331.
[30] Wang N, Yeung D Y. Learning a deep compact image representation for visual tracking[C]//Advances in neural information processing systems. 2013: 809-817.
[31] Wang N, Song Y, Ma C, et al. Unsupervised Deep Tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 1308-1317.
[32] Olshausen B A, Anderson C H, Van Essen D C. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information[J]. Journal of Neuroscience, 1993, 13(11): 4700-4719.
[33] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.
[34] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in neural information processing systems. 2015:

2017-2025.
[35] Wang F, Jiang M, Qian C, et al. Residual attention network for image classification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 3156-3164.
[36] Du W, Wang Y, Qiao Y. Rpan: An end-to-end recurrent pose-attention network for action recognition in videos[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 3725-3734.
[37] Wang Q, Teng Z, Xing J, et al. Learning attentions: residual attentional siamese network for high performance online visual tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4854-4863.
[38] Choi J, Jin Chang H, Yun S, et al. Attentional correlation filter network for adaptive visual tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4807-4816.
[39] Wang G, Luo C, Xiong Z, et al. SPM-Tracker: Series-Parallel Matching for Real-Time Visual Object Tracking[J]. arXiv preprint arXiv:1904.04452, 2019.
[40] Fan H, Ling H. Siamese cascaded region proposal networks for real-time visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019: 7952-7961.
[41] 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.
[42] Lukežič A, Vojíř T, Čehovin L, et al. Discriminative Correlation Filter Tracker with Channel and Spatial Reliability [J][J]. International Journal of Computer Vision (IJCV), 2018, 126(7): 671-688.
[43] Pinheiro P O, Collobert R, Dollár P. Learning to segment object candidates[C]//Advances in Neural Information Processing Systems. 2015: 1990-1998.
[44] Zhang R, Yang W, Peng Z, et al. Progressively diffused networks for semantic visual parsing[J]. Pattern Recognition, 2019, 90: 78-86.
[45] Sun K, Xiao B, Liu D, et al. Deep high-resolution representation learning for human pose estimation[J]. arXiv preprint arXiv:1902.09212, 2019.
[46] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.
[47] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern

recognition. Ieee, 2009: 248-255.
[48] Ren M, Zemel R S. End-to-end instance segmentation with recurrent attention[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 6656-6664.
[49] Arnab A, Torr P H S. Pixelwise instance segmentation with a dynamically instantiated network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 441-450.
[50] Chen H, Qi X, Yu L, et al. DCAN: Deep contour-aware networks for object instance segmentation from histology images[J]. Medical image analysis, 2017, 36: 135-146.