基于孪生神经网络的两阶段目标跟踪方法

张宏伟[1], 李晓霞[1,2], 朱 斌[1], 张 杨[1]

(1. 国防科技大学 电子对抗学院, 安徽 合肥 230037;
2. 脉冲功率激光技术国家重点实验室, 安徽 合肥 230037)

摘 要: 深度学习技术使目标跟踪的精度和鲁棒性得到了很大提高,基于孪生网络的跟踪方法通过在大规模数据集上进行训练,使模型能应对目标的各种形变,缺点是无法排除相似目标的干扰。为此,提出了一种基于孪生网络的两阶段目标跟踪方法。首先,采用修改后的残差网络提取性能更优的深度特征。区域建议网络通过相关滤波调制自适应更新模板,结合时域信息过滤掉易区分的负样本;然后,通过感兴趣池化层提取候选区域固定尺度的特征,并馈送到验证网络进行更精细的分类与回归。为了提升网络对高难度样本的区分能力,采用正负样本对联合训练的方式提高特征匹配的性能。在 OTB100、VOT 标准测试集和 UAV123 无人机航拍数据集上进行了评测,实验结果表明:所提方法能明显改进基准算法的性能。

关键词: 神经网路; 目标跟踪; 区域建议; 相关滤波

中图分类号: TN391.41 **文献标志码:** A **DOI:** 10.3788/IRLA20200491

Two-stage object tracking method based on Siamese neural network

Zhang Hongwei[1], Li Xiaoxia[1,2], Zhu Bin[1], Zhang Yang[1]

(1. School of Electronic Countermeasure, National University of Defense Technology, Hefei 230037, China;
2. State Key Laboratory of Pulsed Power Laser Technology, Hefei 230037, China)

Abstract: Through the introduction of deep learning, the accuracy and robustness of object tracking have been greatly improved. Siamese network based trackers can deal with various deformation of target through training on large-scale datasets, but that makes it difficult to eliminate the interference of similar targets. Therefore, a two-stage tracking method based on Siamese network was proposed. Firstly, the modified residual network was used to extract the deep feature with better performance. Through integrating the temporal information, the template of the region proposal network was adaptively updated through correlation filter modulation, so as to filter out the easily distinguished negative samples. Then, the fixed scale features of candidate regions were extracted by the region-of-interest pooling and fed to the verification network for more refined classification and regression. In order to improve the network's ability to discriminate difficultly distinguished samples, joined training method combining the positive and negative samples was adopted to improve the performance of feature matching. The performance of the proposed method was evaluated on the OTB100, VOT standard benchmarks and the UAV123 aerial benchmark. The experimental results demonstrate that the proposed method can significantly improve the performance of the baseline.

Key words: neural networks; visual tracking; region proposal; correlation filter

收稿日期:2020-12-10; **修订日期:**2021-02-10

基金项目:国家自然科学基金 (61307025)

作者简介:张宏伟,男,博士生,主要从事深度学习、视觉目标跟踪方面的研究。

导师简介:朱斌,男,副教授,博士,主要从事成像侦察与信息处理方面的研究。

0 引 言

目标跟踪广泛应用于智能监控、智能交通、人机交互、机器人导航等民用领域[1−3], 以及远程无人机打击、前视红外、精确制导等军事领域[4−6], 是计算机视觉领域的重要研究方向之一。近年来, 深度学习开始应用于目标跟踪并逐步占据主导地位, 但在面临不可预知的复杂环境时, 如目标遮挡、几何形变、背景扰动、移出视野和快速运动等, 还不能达到很好的效果。

卷积神经网络因学习获得的深度特征具有强大的目标表达能力, 逐渐取代传统的手工特征, 并被引入到目标跟踪任务中取得了很大的发展[7−9]。目前主流的跟踪算法包括相关滤波类和基于深度学习的跟踪方法。相关滤波类[10−14] 利用循环矩阵的性质, 在傅里叶域推导了相关滤波的闭解, 在保证跟踪精度的同时达到了非常快的速度。早期基于深度学习的跟踪方法[15−18] 由于需要对深度网络在线进行微调, 很难达到实时, 例如 MDNet[19] 算法在跟踪的过程中, 针对跟踪目标重建后面的全连接层, 通过在线微调来区分目标和背景, 跟踪速度只能达到 1FPS。在最近两年, 基于孪生网络的深度学习跟踪方法通过在大规模数据集上进行端到端的离线训练, 直接输出目标分类得分和目标框回归结果, 该方法在精度和在线跟踪速度方面的巨大潜力使其成为近两年目标跟踪领域的一个研究热点。

基于孪生网络的跟踪方法中, SINT[20] 将跟踪问题看作是一个分类验证问题, 通过孪生网络对相似性度量函数进行学习, 然后在上一帧的特征空间寻找最相近的区域块。虽然 SINT 采用光流法和自适应采样策略来减少候选目标的数量, 但还是因为候选目标数量庞大, 导致工作速度远远不能满足实时性要求。基于孪生网络结构的 GOTURN[21] 通过全连接层能直接获取目标位置和尺度, 在 GPU 上的运行速度可以达到 100 FPS。由于该模型不能实现模型的在线更新, 算法的鲁棒性与当时的主流算法相比相对要差。SiamFC[22] 在孪生网络的两个分支直接进行卷积运算, 通过响应的最大值确定目标的位置。SiamFC 只能估计目标的中心位置, 而要想对目标的尺寸进行估计, 只有通过多尺度测试来预测尺度的变化, 这种方式不仅增加了计算量, 同时也不够精确。自 SiamFC 提出以来, 出现了一系列基于 SiamFC 的改进算法。DSiam[23] 没有直接采用前一帧的目标特征作为模板, 而是将第一帧的目标特征进行系列变化后使用。RFL[24] 利用卷积长短时记忆网络来动态的生成模板, 但没有显著提升总体跟踪性能。FlowTrack[25] 引入光流来对历史帧进行配准, 利用时空注意力机制融合对齐后的特征, 有效的提高了网络在线应对目标形变的适应能力。SA-Siam[26] 采用语义特征和表观特征互补的形式, 先分别得到各自的响应, 最后再加权求和来估计目标位置。SiamRPN[27] 通过引入区域建议网络来对目标进行分类和回归, 通过锚点结构直接给出目标位置和尺度信息, 端到端的网络设计不需要多尺度预测, 进一步提高了目标跟踪的速度和鲁棒性。DaSiamRPN[28] 利用目标检测结果来扩展训练数据集, 构建语义信息丰富的负样本进行离线训练, 来提高网络的辨别能力, 并利用了历史帧的上下文信息来解决相似干扰因素的影响。

SiamRPN 类跟踪方法在离线阶段通过在大规模数据集上进行训练, 具有很强的泛化能力, 在处理目标形变方面具有明显的优势。但局限性在于虽然对目标的各种形变不敏感, 但对相似性目标的分辨能力较差, 如图 1 所示。SiamRPN 采用的措施是通过加余弦窗和尺度惩罚的方式, 选择离上一帧目标较近的相关响应峰值作为跟踪结果, 这显然与目标跟踪能区分单个目标的任务不相适用。如何保证模型具有很强泛化能力的同时, 提高区分跟踪目标与其他干扰物的能力是需要解决的主要难点问题。

图 1 SiamRPN 分类响应图

Fig.1 Classification response map of SiamRPN

1 文中算法

文中在 SiamRPN 的基础上, 设计了一种基于孪生网络的两阶段跟踪方法, 主要解决一阶段的孪生网络为适应目标各种形变无法兼顾对相似性目标的区分, 网络结构如图 2 所示。网络的输入为参考图像 z和以上一帧目标为中心的小范围搜索区域x, 两者经

共享权值的主干网络提取对应的深度特征 $f_\theta(z)$和 $f_\theta(x)$, 主干网络采用了更深的残差网络 Resnet50; $f_\theta(z)$经相关滤波 (Correlation Filter, CF) 调制得到目标模板w,模板通过滑动平均的方式进行自适应更新, 以综合时域信息提高目标的显著性。接着将模板w作为卷积核与$f_\theta(x)$进行卷积运算, 计算得到的相关矩阵输入到区域建议网络 (Region Proposal Network, RPN) 完成背景和目标的初步分类; 最后, 使用感兴趣池化层 (RoI Pooling) 将候选区域转化为固定尺度的特征, 与模板特征一起进行更精细化的分类与回归。

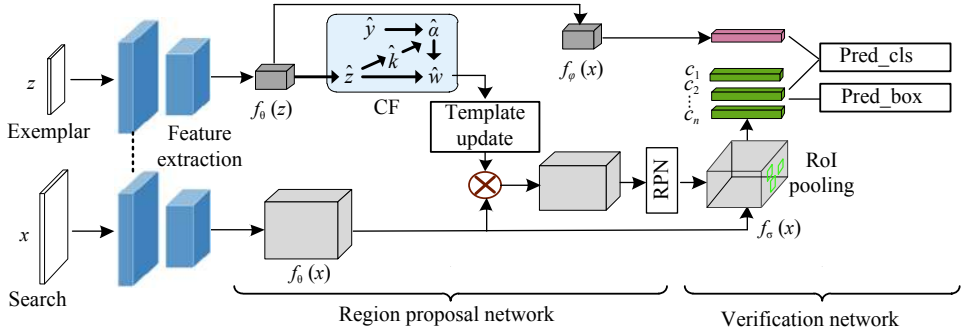

图 2 基于孪生网络的两阶段跟踪方法结构示意图

Fig.2 Schematic diagram of the two-stage tracking method based on Siamese network

1.1 深度特征提取

SiamRPN 算法的主干网络采用的是 AlexNet 网络, 使用 conv5 层的特征对目标进行不同尺度和长宽比的预测, 不能提供多元化的特征表示。一般来说, 深度网络不同层具有不同的意义, 这里采用更深的 ResNet50 网络进行实验, 分析提取不同层次的特征对于模型性能的影响。考虑到感受野的变化, 深层特征具有丰富的语义信息, 在运动模糊、大形变等挑战场景中更有利于定位。

为了将更深的网路 ResNet50 运用于孪生网络跟踪, 需要对模型的结构进行一下调整: 在原来的 ResNet50 中, 步长为 32, 特征尺度太小不适合相关运算, 通过修改 P3 和 P4 模块, 取消了下采样操作, 代替以空洞卷积层来扩大其感受野; 因为多次叠加空洞卷积会损失信息的连续性, 因此空洞卷积只使用在残差单元中第二个卷积层。

为了分析网络的深度对 SiamRPN 跟踪性能的影响, 提取了 ResNet50 不同深度的特征层输入到区域建议网络进行训练。利用 OTB 数据集对训练好的多个模型进行了测试, 结果表明采用 AlexNet 网络模型的跟踪精度为 0.805, 采用 ResNet50 网络的 P2、P3 和 P4 模块输出进行预测的跟踪精度分别为 0.823、0.830、0.813, 可以看到采用 P3 特征层的模型精度最高, 说明并不是深度越深精度越高。因此特征提取模块采用了调整后的 Resnet50 网络, 提取 P3 特征层进行区域建议, 使模型实现更优的性能。图 3 分别为采用不同基础网络提取的特征可视化, 可以看到修改后 ResNe50 网络提取的特征质量更高, 具有更好的应对背景干扰的能力。

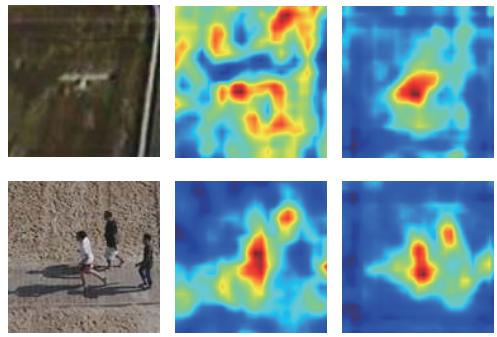

图 3 不同网络提取的特征

Fig.3 Features extracted by different networks

1.2 相关滤波调制

为了提高区域建议网络提取候选目标的质量, 改进方法是通过相关滤波调制在 RPN 阶段对模板进行自适应更新, 具体过程如图 2 中的 CF 模块。对于获取的目标特征先与相关滤波算子相乘, 计算获得的目标模板再与搜索区域特征做互相关运算得到高斯形状的相关响应, 经相关滤波之后的互相关响应可以表示为:

$$R_{\theta,s,b}(z,x)=s\omega(f_\theta(z))\otimes f_\theta(x)+b \tag{1}$$

相关滤波模板$w\in\mathbb{R}^{m\times m}$的求解通过最小化目标函数得到，并通过时域到频域的转化提高矩阵的运算速度。对于给定的训练样本$Z\in\mathbb{R}^{d\times n}$，该优化问题可以描述为：

$$\arg\min_w \frac{1}{2n}\left\|Z^{\mathrm{T}}w-y\right\|^2+\frac{\lambda}{2}\|w\|^2 \tag{2}$$

式中：$y\in\mathbb{R}^n$和λ分别表示目标函数和正则化系数。将$Z^{\mathrm{T}}w-y$用r表示，采用拉格朗日乘数法求解优化问题[29]：

$$L(w,r,v)=\frac{1}{2n}\|r\|^2+\frac{\lambda}{2}\|w\|^2+v^{\mathrm{T}}(r-Z^{\mathrm{T}}w+y) \tag{3}$$

式中：v为拉格朗日乘子，将该式进行求导运算，并令偏导数为 0，得到目标模板w的表达式为：

$$\begin{cases}\hat{k}=n^{-1}(\hat{z}^*\odot\hat{z})+\lambda\\ \hat{\alpha}=n^{-1}\hat{k}^{-1}\odot\hat{y}\\ \hat{w}=\hat{\alpha}^*\odot\hat{z}\end{cases} \tag{4}$$

式中：$\wedge$表示傅里叶变换；$\odot$表示矩阵的点乘。通过变换到傅里叶域后，相关滤波的计算变为元素点乘的形式，可以加快运算速度。相关滤波调制的优点是在跟踪的过程中，模型可以综合历史帧进行自适应更新，提高跟踪目标的显著性，模板更新采用滑动平均的更新策略：

$$\hat{w}=\eta\hat{w}_n+(1-\eta)\hat{w}_{n-1} \tag{5}$$

式中：η为目标模板学习系数，实验中取 0.9。为了进行对比分析，这里对原始网络和使用相关滤波调制后的网络在 GOT-10k 同一个数据集上进行训练，通过提取网络中间层的分类响应进行比较。

如图 4 所示，第一行为原始网络的区域建议结果，第二行为改进之后的网络输出结果。从响应图对比可以看到，通过相关滤波调制之后，目标候选框取样更多的集中到了目标附近。改进之后的网络可以提高目标的显著性，在 RPN 阶段过滤掉易区分的负样本，像第一组序列中右下角的摩托车、第 2 组序列中三轮车附近的干扰车辆。更加准确的候选框提取有助于减轻第二阶段验证网络的分类压力，降低虚警率。

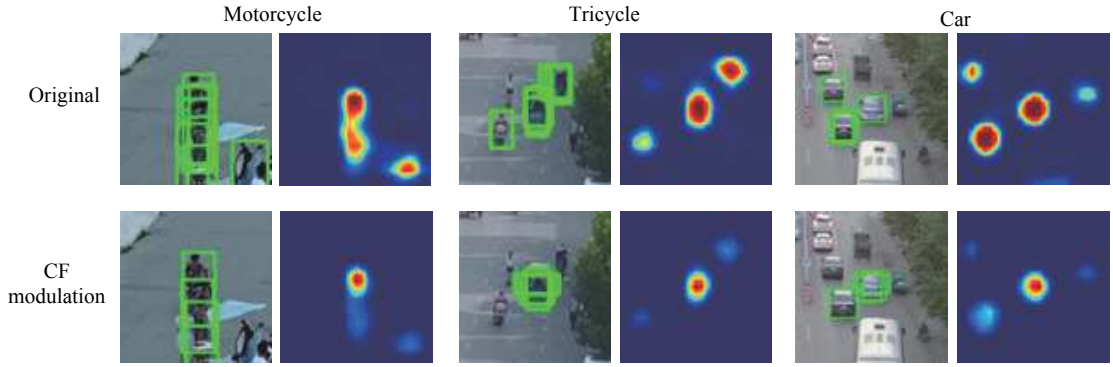

图 4 使用相关滤波调制后的网络输出对比分析

Fig.4 Comparative analysis of network outputs after applying correlation filter modulation

1.3 验证网络

泛化能力强会带来高的误检率，这就是一阶段孪生网络 SiamRPN 无法很好的将目标与相似干扰物区分开来的原因。前面针对区域建议网络增加了相关滤波调制来提高目标的显著性，但还不能完全解决高难度样本的区分，因此设计了区域建议加精细验证的两阶段网络结构。

如图 5 所示，区域建议网络获得的多个候选区域利用感兴趣池化层 (Region of Interest pooling, RoI poooling) 提取候选框各自的特征，馈送到验证网络进行进步的分类和回归。对于分类分支，提取的候选区域特征先通过 2 个 DBL 卷积组对特征进行优化，再经池化层和后续的全连接层得到分类向量，最后与模板的特征向量进行相似性度量。对于回归分支，候选目标特征先与模板特征进行相关运算，再进行目标框的回归。目标跟踪的核心任务要求模型能够区分单个个体，因此设计的验证网络需要在训练过程中使得模型输出的特征具有较强的分辨力，即满足目标与跟踪过程中各个状态之间的距离要小于与干扰物之间的距离。

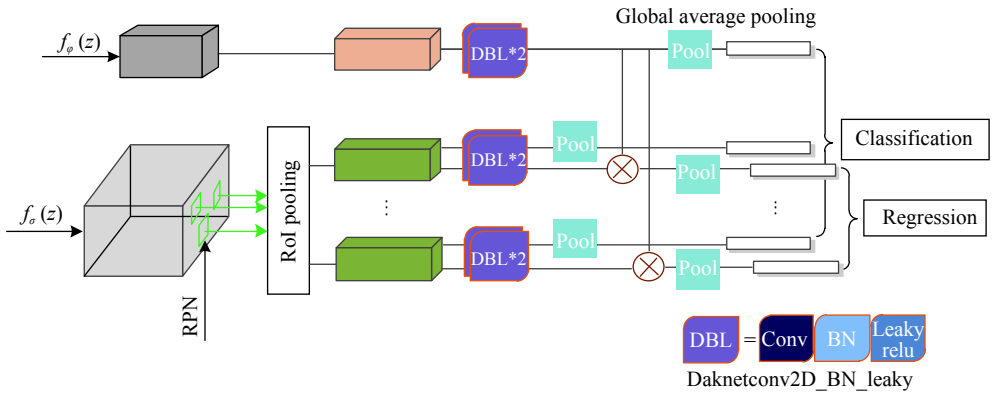

图 5 验证网络结构示意图

Fig.5 Schematic diagram of the verification network

对于目标特征$z \in \mathbb{R}^{k\times k\times c}$和第$i$个候选框$c_i \in \mathbb{R}^{k\times k\times c}$，为了增加分类的辨别力，将正负样本组成样本对$T=\{\langle c_i,c_j,z\rangle\}$来训练损失函数，其中$\langle c_i,z\rangle$来自同一目标的样本特征，$\langle c_j,z\rangle$来自不同目标的样本特征，对于$N$个集合，分类损失可以表示为：

$$L_{cls}=\frac{1}{N}\sum_{\langle c_i,c_j,z\rangle\in T}\max(d_i-d_j+m,0) \tag{6}$$

式中：d为目标与样本的距离；$m>0$为预设边际值。对于回归分支，采用跟踪模板引导目标框回归的形式，首先对候选框特征进行调制来编码相关运算：

$$\bar{x}_i=h_{out}(h_x(c_i)\odot h_z(z)) \tag{7}$$

式中：$\odot$表示哈德曼乘积；h_x和h_z为特征映射，采用3×3的卷积；h_{out}是为了保证得到需要的特征维度，采用1×1的卷积。网络训练采用Smoth L1损失：

$$L_{loc}=\frac{1}{N_{Prop}}\sum_i s_i^* L_{smoth}(p_i,q_i) \tag{8}$$

式中：p_i为估计的中心位置和尺度偏差；q_i为标签；s_i^*表示第i个候选框是否判断为目标。

通过正负样本对联合训练的方式可以大大提高模型的分类本领，图 6 为训练 10 代之后的模型在跟踪 UAV123-group1 过程中对不同样本的分类结果，目标周围同时存在其他人不定时的交叉影响。虚线为第一帧目标与后面各状态之间的相似性度量，实线为目标与周围干扰项之间的相似性度量，度量函数采用余弦距离。可以看到在整个跟踪过程中，验证网络对同类物体具有很好的分辨力，弥补了 SiamRPN 分辨能力差的缺点。

图 7 选取了多个样本组进行进步的分析，从它们

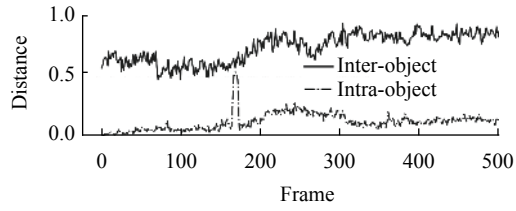

图 6 类间与类内分类距离对比

Fig.6 Comparison of classification distances of the inter-object and intra-object

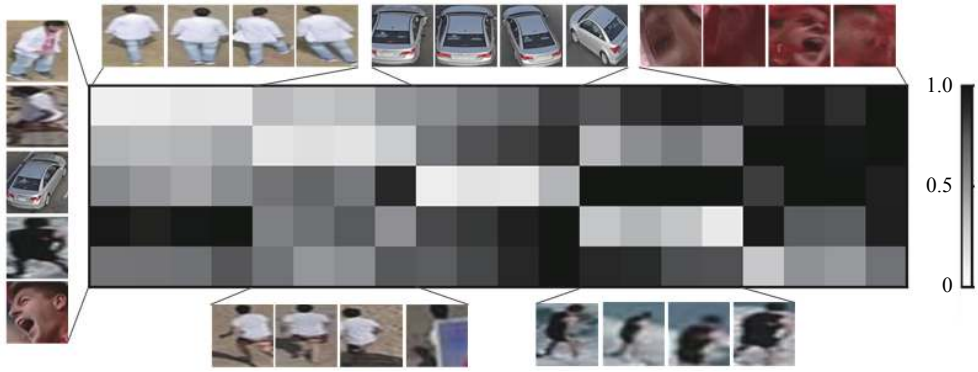

图 7 不同取样组分类距离对比结果

Fig.7 Comparison of classification distance between different sampling groups

的对比结果可以看出,同一个体的多个取样间余弦距离都较小,不同个体的取样之间距离相对较大。在目标发生部分遮挡与剧烈形变的情况下,分类距离有所上升,但依然可以将之与干扰物区分开来。

1.4 算法整体流程

综合上述对文中算法关键部分的描述,主要跟踪步骤如下所示。

输入:视频序列 $I_1, I_2, \cdots, I_n$,初始目标位置 $p_0=(x_0,y_0)$,初始目标尺寸 $s_0=(w_0,h_0)$

输出:每一帧目标位置 $p_t=(x_t,y_t)$和目标尺寸 $s_t=(w_t,h_t)$

For n=1, 2, …, n, do:

if n=1:

1.1 裁剪目标小块区域 z,利用主干网络提取目标的深度特征 $f_\theta(z)$;

1.2 根据公式 (4) 对 $f_\theta(z)$进行相关滤波调制,并初始化目标模板 $w=w_0$;

if n>1:

2.1 以上一帧目标中心位置裁剪第 n 帧的搜索区域,提取搜索区域的特征 $f_\theta(x)$;

2.2 将 w与 $f_\theta(x)$进行相关运算,相关结果输入到区域建议网络,通过非极大值抑制得到多个候选目标;

2.3 利用 RoI pooing 层提取候选目标固定尺度的特征 $\{c_i\}$;

2.4 $f_\theta(z)$通过卷积运算调整尺度大小得到映射矩阵 $f_\varphi(z)$, $f_\varphi(z)$与 $\{c_i\}$进行相似性度量得到最终目标估计,对应的特征 c_t依照公式 (9) 进行相关调制后用于目标框回归;

2.5 根据跟踪结果裁剪 $f_\theta(x)$对应的目标区域,并计算当前帧的模板 w_n,根据公式 (5) 对模板 w进行更新;

Until 视频序列结束

2 实验与结果分析

2.1 实验参数设置

主干网络采用改进后的 ResNet50 网络,提取第 3 个残差模块的特征,先在 ImageNet 数据集上预训练,再采用迁移学习的方式利用 GOK-10k、YouTube-BB 数据集进行微调。区域建议网络和验证网络采用交替训练的方式,训练代数为 20 代,学习率从 10^{-3}~10^{-5} 依次递减。模板和搜索区域输入尺寸分别为 127×27 和 255×255。采用随机梯度下降 (Stochastic Gradient Descent, SGD) 优化算法对网络进行训练,参数更新时的动量 m 取 0.9,权值衰减 γ取 0.000 5。网络构建采用 PyTorch 深度学习框架,实验平台:CPU 为 Intel Xeon E5-2650@2.20 GHz, GPU 为 NVIDIA TITAN V。

2.2 利用 OTB100 测试集进行评测

OTB100 标准测试集共计 100 个测试序列,视频平均长度为 589 帧。根据跟踪的难点问题,对每个视频定义了属性标签,包括:遮挡、旋转、变形、光照变化和尺度改变等。评价指标包括准确率图和成功率图,其中准确率图 (Precision Plot) 基于中心定位误差 (Center Location Error, CLE) 对跟踪算法进行评价。CLE 定义为跟踪算法预测的目标框中心与标注框中心之间的距离,统计出距离小于一定阈值的图片数目,其占视频总帧数的百分比表示该阈值下的准确率。一般情况下采用阈值为 20 pixel 对应的准确率作为该项评测准则的具体指标。成功率图 (Success Plot) 采用重叠率 (Intersection Over Union, IOU) 作为评价的基准,IoU 指的是跟踪算法预测的目标框 A_t与标注框 A_{gt}之间的交并比,即 $\phi_t=|A_t\cap A_{gt}|/|A_t\cup A_{gt}|$。统计出 ϕ_t大于一定阈值下的跟踪图片数目占视频总帧数的百分比,即为该阈值下的成功率。该评测准则的具体指标为 AUC(Area Under Curve),即成功率曲线与横轴围成的面积。

OTB100 标准测试集包含了 29 种算法的评价结果,这里将文中算法与排名前 9 的算法 TLD、OAB、CSK、ASLA、Struck、DSST、KCF、SAMF、MEEM,以及开源的跟踪算法 MUSTER、SRDCF、SiamFC、CFNet、SiamRPN 和 ECO 进行了对比试验。实验采用 OPE(One Pass Evaluation) 测试方式,只利用初始帧的目标信息,然后跟踪完整个测试序列,测试结果如图 8 所示。

从精确度曲线和成功率曲线来看,文中算法的精确度与成功能率比 ECO 高出了 6.4% 和 4.0%。与原来的基准算法 SiamRPN 相比提升效果明显,精确度与成功率分别提高了 12.3% 和 9.7%。性能提升的原因主要在于二阶段网络结构设计,SiamRPN 简单的将

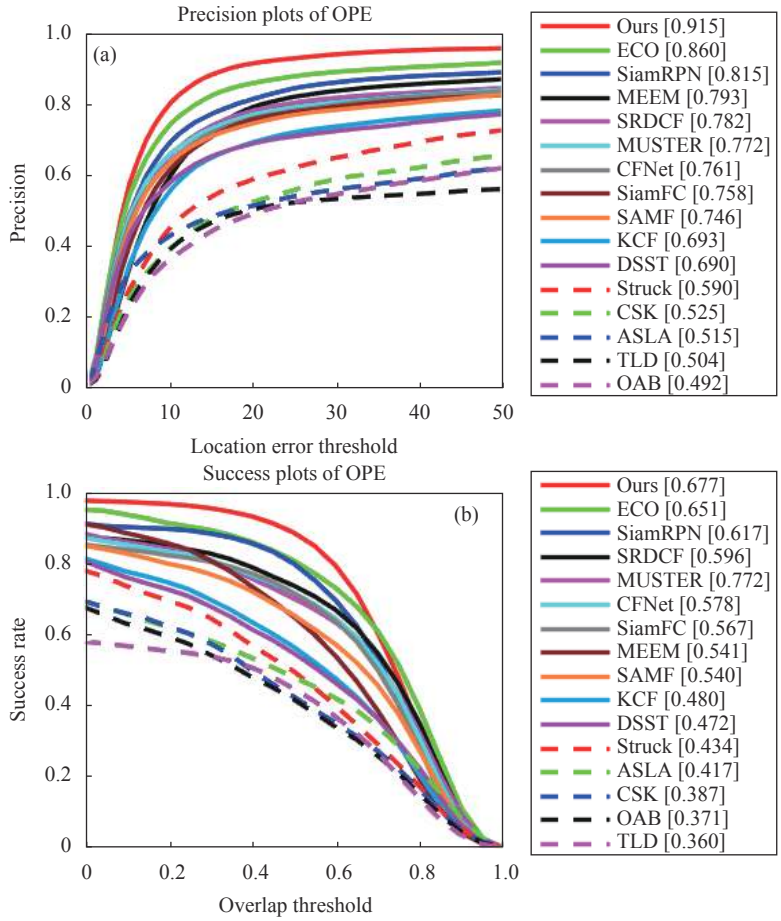

图 8 OTB100 标准测试集评价: 精确度曲线和成功率曲线

Fig.8 Evaluation on the OTB100 benchmark: Precision plot and success plot

相关响应最高的位置作为跟踪结果, 而文中提出的网络首先结合时域信息对易区分的样本进行滤除, 再利用区域建议加验证网络的方式对高难度样本进行区分, 使网络具有更好的分辨力。

2.3 利用 VOT 测试集进行评测

VOT 标准测试集包含了 60 个更具挑战性的视频, 有自己比较系统的评价体系, 采用准确率 (Accuracy)、鲁棒性 (Robustness) 和期望平均重叠率 (Expected Average Overlap, EAO) 3 项指标对跟踪算法进行评价。Accuracy 用来评价跟踪算法的准确度, 定义为跟踪过程中有效帧的平均重叠率, 即 $\rho_A = 1/N_{N_{\text{valid}}} \sum_{i=1:N_{\text{valid}}} \phi_i$, 其中 N_{valid} 表示为有效帧的数量。Robustness 用来评价跟踪算法的稳定性, 定义为跟踪失败的次数。EAO 指标将所有测试视频的长度考虑其中, 其计算方法如下: 算法对一个视频进行跟踪, 跟踪失败后会进行重新初始化, 这样一个视频被分割为不同长度的片段。对于一个长度为 N_S 的片段平均重叠率为 $\Phi_{N_s} = 1/N_s \sum_{i=1:N_s} \phi_i$, 对多个不同长度的序列进行跟踪, 计算视频长度在区间 $[N_{lo}, N_{hi}]$ 的期望平均重叠率为 $\hat{\Phi} = 1/(N_{hi} - N_{lo}) \sum_{N_s=N_{lo}:N_{hi}} \Phi_{N_s}$。测试方式分为基本测试和非监督测试, 不同之处在于在基本测试条件下, 当跟踪失败后, 会重新用标注框对跟踪器进行初始化。图 9 为 AR 得分, 综合了精确度与鲁棒性两项指标对算法进行排序, 越靠右上表示性能越好。图 10 为非监督条件下的重叠率曲线, 采用 AUC 作为算法排名的评价指标。

从表 1 的测试集评价结果 (评价指标包括基本测试条件下的精确度、鲁棒性, 预期平均重叠率以及非监督条件下的平均重叠率、速度) 来看, 在基本测试条件下, 所提算法的 A/R 得分为 0.601 1/14.515 9, 与一阶段孪生网络 SiamRPN 相比, 在鲁棒性方面的性能提升更加明显, 结合前面的理论分析可知文中从多

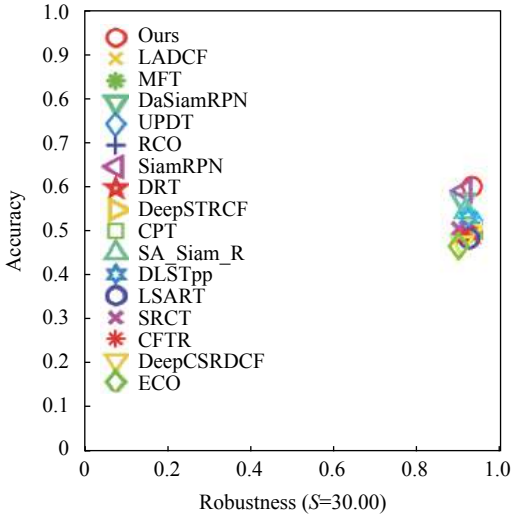

图 9 基本测试条件下 AR 得分

Fig.9 AR scores in baseline test condition

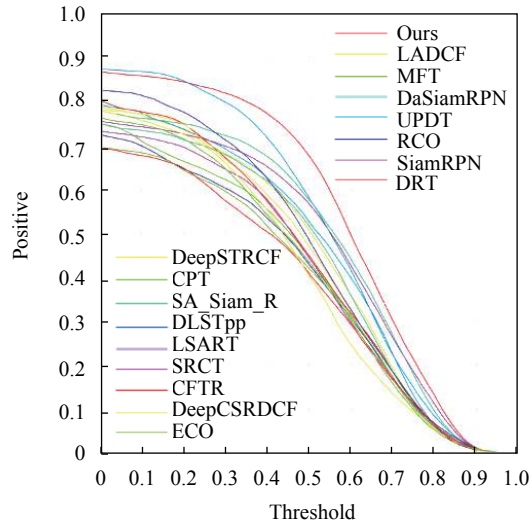

图 10 非监督条件下重叠率曲线

Fig.10 Overlap rate curves in unsupervised condition

表 1 VOT 标准测试集评价结果

Tab.1 Evaluation results on the VOT benchmark

	Baseline			Unsupervised		
	A-R rank		EAO	Overlap	Speed	
	Overlap	Failures	EAO	AUC	Normalized	FPS
Ours	0.601 1	14.515 9	0.383 3	0.533 9	3.496 1	20.245 1
LADCF	0.491 1	9.925 3	0.381 1	0.418 2	0.123 0	0.557 3
MFT	0.491 9	10.766 2	0.379 4	0.391 7	0.194 5	0.623 2
DaSiamRPN	0.569 1	18.441 5	0.378 5	0.468 4	17.818 3	64.414 3
UPDT	0.515 4	11.417 2	0.371 9	0.444 4	0.088 4	0.469 7
RCO	0.498 9	10.700 4	0.371 1	0.383 0	0.204 6	0.720 3
SiamRPN	0.591 5	19.632 5	0.369 1	0.456 8	20.342 6	86.784 3
DRT	0.495 8	13.947 6	0.349 0	0.419 1	0.123 7	0.456 8
DeepSTRCF	0.506 2	14.548 6	0.338 3	0.433 3	0.560 5	3.114 4
CPT	0.488 8	16.620 7	0.332 1	0.375 7	0.877 1	5.184 2
SA_Siam_R	0.544 4	16.403 0	0.331 1	0.425 0	6.776 1	32.364 4
DLSTpp	0.529 7	14.937 4	0.321 3	0.497 8	1.293 0	8.175 9

方面提高孪生网络对各种干扰的辨别能力，使得失败次数大大下降。EAO 与 AUC 指标分别比排名第 2 的 LADCF 和 DLSTpp 提高了 0.6%、3.6%，说明算法在非监督测试条件的优势体现得更为明显。由于采用更复杂的网络，跟踪速度为 20.2451FPS，比 SiamRPN 有所下降，但仍然优于排名靠前的 LADCF、MFT 算法。

2.4 利用 UAV123 航拍数据集进行测试

为了进一步测试文中算法性能，采用无人机航拍数据集 UAV123 对算法进行了测试。数据集使用专业级无人机 (DJI S1000) 进行拍摄，相机固定在可控框架系统 (DJI Zenmuse Z15) 上，跟踪高度在 5~25 m 之间。相机采用焦距为 12 mm 的松下 GH4，视频序列以 30~96 帧/s 的帧速和 720 p~4 k 的分辨率记录。UAV123 数据集包含 123 个视频序列，总帧数达 110 000 帧，属于长时间跟踪数据集。跟踪目标包括汽车、卡车、船只、人和空中无人机，以顶视角的方式进行拍摄。在该数据集中的评价采用了同 OTB100 测试集一样的评价标准，利用精确度图和成功率图对算法性能进行分析。

无人机数据集与 OTB 数据集相比，面临的情况更加复杂，跟踪视频的长度普遍较长，相机晃动以及目标频繁移出视野对目标跟踪带来了更大挑战。从图 11 的评测结果来看，SiamRPN 因为对目标形变具有较好的适应能力，在长时间跟踪数据集上的性能要优于 ECO 算法，这与在 OTB 测试集上的评分正好相反。文中算法与 SiamRPN 相比准确度与成功率分别提高了 5.4% 和 7.4%，在没有采用长时间跟踪策略的情况下，得益于两阶段跟踪网络更准确的分类和更高的回归精度，文中方法在无人机航拍数据集上也有很好的表现。

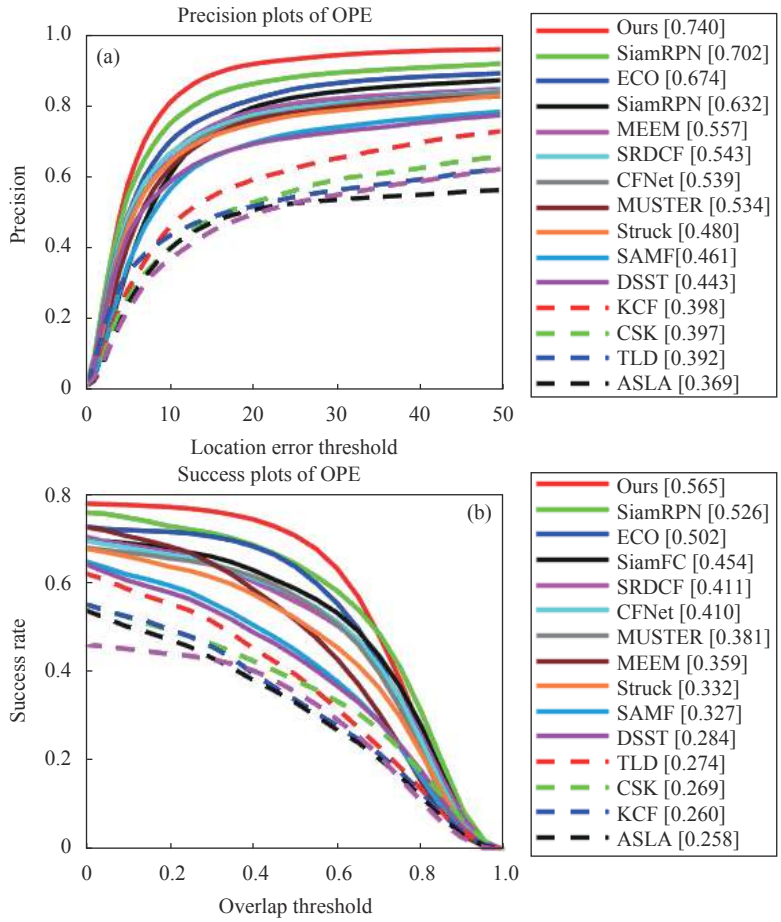

图 11 UAV123 数据集测试结果: 精确度曲线和成功率曲线

Fig.11 Test results on the UAV123 dataset: Precision plot and success plot

2.5 定性分析

为了对算法的性能有个定性的分析，列举了 10 种算法针对多种跟踪场景的对比结果，这些跟踪场景包含多种挑战性因素: 剧烈形变、相似干扰物交叉影响、目标遮挡、背景干扰等，如图 12 所示 (OTB-Diving、 OTB-Skating2、 UAV123-bike2、 UAV123-group1),在序列 OTB-Diving 中，运动员身体完全变形且具有较快的运动速度，这时候基于区域建议的 SiamRPN 与文中方法的优势体现出来，能较好的适应目标形变，同时文中方法采用两次目标框回归，回归精度有了进步的提升。在序列 OTB-Skating2 中，被跟踪目标男运动员与干扰物属于同一类别，且频繁交叉遮挡造成干扰，SiamRPN 由于缺乏区分同类目标的能力，常常错误将女运动员作为跟踪结果。而文中方法经过 RPN 阶段相关滤波调制减少错误样本和第二阶段验证网络更加精准分类，具有更好的鲁棒性。相类似情况还有序列 UAV123-group1，在目标与干扰物相互干扰的过程中，文中方法都能准确的跟踪目标。在序列 UAV123-bike2 小目标的跟踪过程中，目标存在多次遮挡的情况，其他算法最后都跟丢目标。得益于

强的分辨本领，文中方法在抗遮挡方面也有很好的表现。在无人机空对地的跟踪场景中，相似干扰与频繁遮挡是需要解决的主要挑战，文中方法在该方面体现出了更好的适用性。

两阶段孪生网络性能的提升主要来源于两方面：一方面是通过相关滤波调制以及验证网络相似性度量提升了网络的分类能力；另一方面通过两次目标框回归，回归的准确性也进步提高。图 13 显示了最终网络对区域建议结果的分类得分，可以看出，算法对错误采样以及相似性目标有很好的区分度，成功地

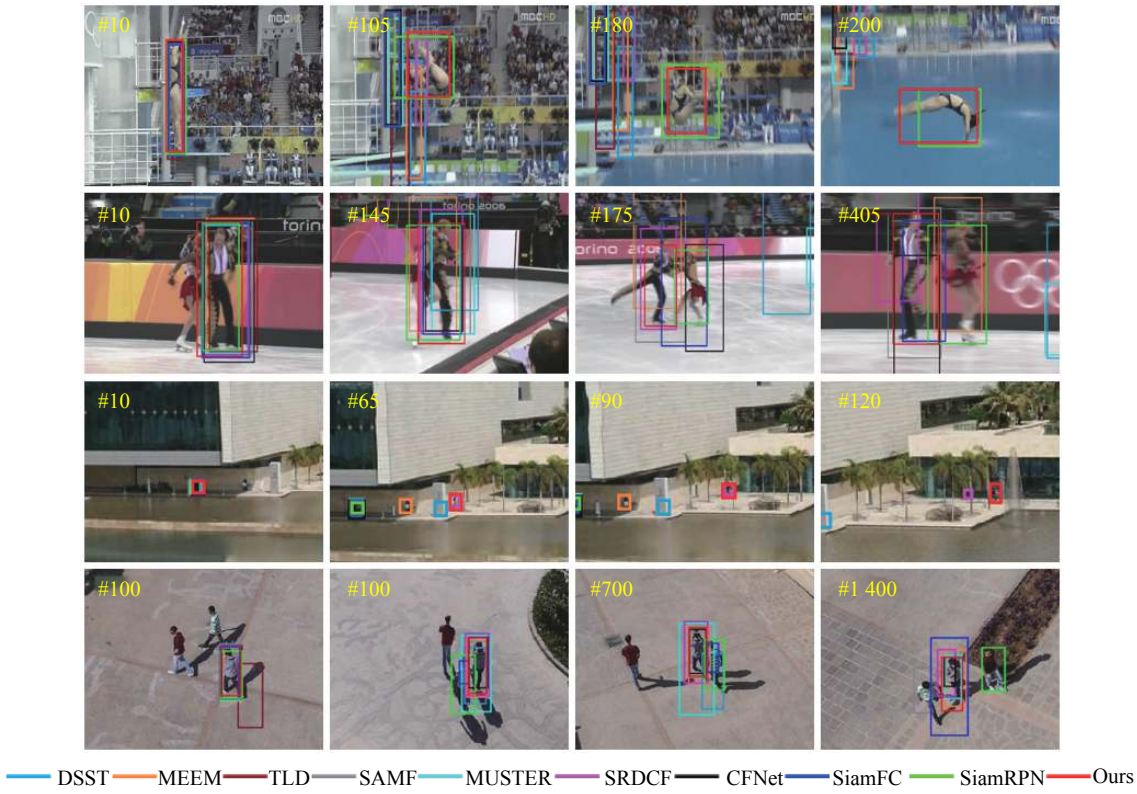

图 12 针对挑战性视频的定性评价

Fig.12 Qualitative evaluation of challenging sequences

图 13 两阶段孪生网络应对相似干扰的能力

Fig.13 Ability to deal with the similar interference of the two-stage Siamese network

解决了原 SiamRPN 网络无法处理相似目标干扰的问题。

图 14 为改进网络回归精度的对比，从图中可以看出，通过两次目标框回归，跟踪精度有了更进一步的提升。例如在序列 UAV123-car1 跟踪到第 401 帧的时候，另外一辆汽车由于靠得太近，原 SiamRPN 算法将黑色车辆也包围进了跟踪框，而两阶段网络能够准确识别目标与干扰。另外，在对小尺度目标 (序列 UAV123-uav3)、非整体目标 (序列 UAV123-bike1) 的测试也体现了两阶段网络的优势。

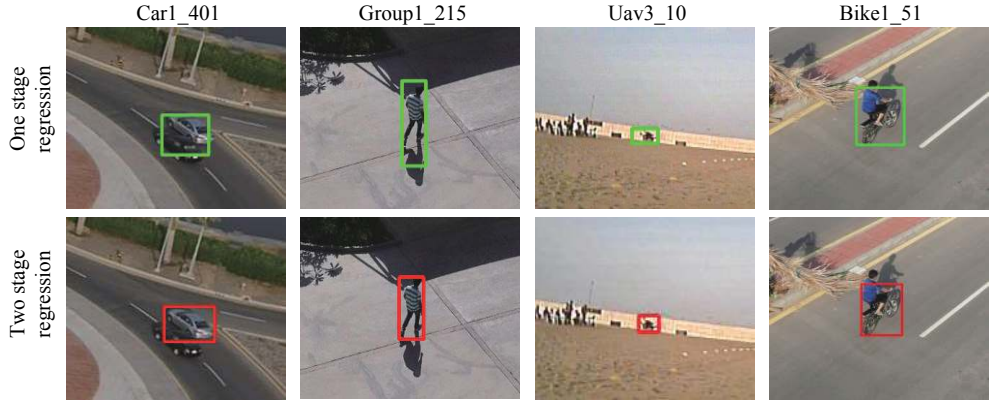

图 14 两阶段孪生网络回归精度对比

Fig.14 Comparison of regression precision of the two-stage Siamese network

图 15 为各模块运行时间的对比，测试序列为 UAV123-car1，采用 GPU 加速，硬件参数见第 2.1 节。从图中可以看到，特征提取模块占用了大部分的计算资源，第二阶段验证网络的运行时间比区域建议模块稍长。在有 GPU 加速的情况下，两阶段孪生网络依然能达到实时。在无人机平台等嵌入式设备不具备 GPU 运行条件下，需依靠地面端通过数据链传输的方式来实现。

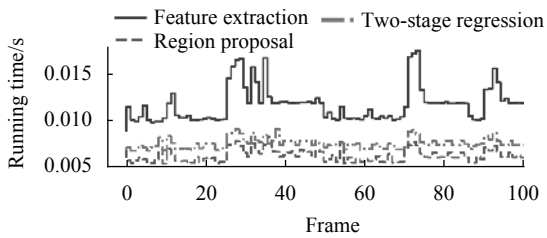

图 15 各模块运行时间测试

Fig.15 Running time test of each module

3 总结与展望

文中在孪生网络的基础上，提出了一种基于孪生网络的两阶段跟踪方法。在 RPN 阶段，通过相关滤波调制和锚点结构设计，得到初步的目标候选框，获取的候选框经感兴趣池化层提取特征后输入到验证网络进行更精准化的分类和回归。相对于一阶段 SiamRPN 孪生网络，文中方法较好的解决了原来算法无法兼顾泛化能力与抗干扰性的问题，相关滤波调制加上两次目标框回归使模型具有更好的精确度。在多个标准测试集上的评测表明，文中方法在保证较快跟踪速度的前提下，跟踪精度与区分相似干扰物的能力大大提升。由于缺少长时间跟踪策略，模型在跟踪失败后无法对全图进行目标搜索，值得更进一步的研究。

参考文献：

[1] Smeulder A, Chu D, Cucchiara R, et al. Viusal tracking: An experimental survey [J]. ***IEEE Transactions on Pattern Analysis and Machine Intelligence***, 2014, 36(7): 1442-1468.

[2] Hou Z Q, Han C X. A survey of visual tracking [J]. ***Acta Automatica Sinica***, 2016, 32(4): 603-607. (in Chinese)

[3] Wang N, Shi J, Yeung D-Y, et al. Understanding and diagnosing visual tracking systems[C]//IEEE International Conference on Computer Vision, 2015: 3101-3109.

[4] Qi He. Research on target tracking and key techniques of electro-optical image guidance system[D]. Beijing: Beijing Institute of Technology, 2016: 1-4. (in Chinese)

[5] Yang Chunwei, Wang Shicheng, Liao Shouyi, et al. Forward-looking-infrared building object tracking based on sparse representation of covariance descriptor [J]. ***Infrared***

Technology, 2016, 38(5): 389-395. (in Chinese)

[6] Hossain S, Lee D J. Deep learning-based real-time multiple-object detection and tracking from aerial imagery via a flying robot with GPU-Based embedded devices [J]. ***Sensors***, 2019, 19(15): 1-2.

[7] Qiu Z L, Zha Y F, Zhu P, et al. Visual tracking algorithm based on online feature discrimination with Siamese network [J]. ***Acta Optica Sinica***, 2019, 39(9): 2247.

[8] Li Yong, Yang Dedong, Han Yajun, et al. Siamese neural networks object tracking integrating [J]. ***Acta Optica Sinica***, 2020, 40(4): 0415002. (in Chinese)

[9] Shi Guoqiang, Zhao Xia. Object tracking algorithm based on jointly-optimized strong-coupled Siamese region proposal network [J]. ***Journal of Computer Applications***, 2020, 40(10): 2822-2830. (in Chinese)

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

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

[12] Henriques J, Caserio R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]//European Conference on Computer Vision, 2012: 702-715.

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

[14] Valmadre J, Bertinetto L, Henriques J F, et al. End-to-end representation learning for correlation filter based tracking[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2805-2813.

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

[16] Zhang K, Liu Q, Wu Y, et al. Robust visual tracking via convolutional networks without training[C]//IEEE Transactions on Image Processing, 2015: 1779-1792.

[17] Wang L, Ouyang W, Wang X, et al. Visual tracking with fully convolutional networks[C]//IEEE International Conference on Computer Vision, 2015: 3119-3127.

[18] Ma C, Huang J, Yang X, et al. Hierarchical convolutional features for visual tracking[C]//IEEE International Conference on Computer Vision, 2015: 3074-3082.

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

[20] Tao R, Gavves E, Smeulders A W. Siamese instance search for tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2016: 903-909.

[21] Held D, Thrun S, Savarese S. Learning to track at 100 fps with deep regression networks[C]//European Conference on Computer Vision, 2016: 749-765.

[22] Bertinetto L, Valmadre J, Henriques J F, et al. Fully convolutional siamese networks for object tracking[C]// Proceedings of the European Conference on Computer Vision Workshop, 2016: 850-865.

[23] Guo Q, Feng W, Zhou C, et al. Learning dynamic Siamese network for visual object tracking[C]//IEEE International Conference on Computer Vision, 2017: 1781-1789.

[24] Yang T, Chan A B. Recurrent filter learning for visual tracking[C]//IEEE International Conference on Computer Vision Workshops, 2018: 2010-2019.

[25] Zhu Z, Wu W, Zou W, et al. End-to-end flow correlation tracking with spatial-temporal attention[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 548-557.

[26] He A, Luo C, Tian X, et al. A twofold Siamese network for real-time object tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4834-4843.

[27] Li B, Yan J, Wu W, et al. High performance visual tracking with siamese region proposal network[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8971-8980.

[28] Zhu Z, Wang Q, Li B, et al. Distractor-aware siamese networks for visual object tracking[C]//IEEE European Conference on Computer Vision, 2018: 103-119.

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