融合残差连接与通道注意力机制的 Siamese 目标跟踪算法

邵江南[1,2], 葛洪伟[1,2)*]

[1)] (江南大学人工智能与计算机学院 无锡 214122)
[2)] (江南大学江苏省模式识别与计算智能工程实验室 无锡 214122)
(ghw8601@163.com)

摘 要: 针对 Siamese 跟踪算法在目标形变、相似物体干扰等复杂情况下容易跟踪漂移或丢失的问题, 提出一种融合残差连接与通道注意力机制的目标跟踪算法. 首先, 通过残差连接将模板分支网络提取的浅层结构特征与深层语义特征进行有效的融合, 以提高模型的表征能力; 其次, 引入通道注意力模块, 使模型自适应地对不同语义目标特征通道加权, 以提高模型的泛化能力; 最后设计并提出一种基于相关性响应值的权重掩码, 在离线训练时提高相似语义目标损失值的权重, 使模型在端到端的离线学习中增强对相似语义目标的辨别力. 在标准跟踪数据集 OTB, TempleColor128, VOT2016 和 VOT2018 上与主流跟踪算法进行对比实验, 结果表明, 该算法在跟踪精度和成功率上都展现了极强的竞争力, 具有优越的实时性和可靠性.

关键词: 目标跟踪; 卷积神经网络; 孪生网络; 特征融合; 通道注意力机制
中图法分类号: TP391.41 **DOI:** 10.3724/SP.J.1089.2021.18340

Siamese Object Tracking Algorithm Combining Residual Connection and Channel Attention Mechanism

Shao Jiangnan[1,2] and Ge Hongwei[1,2)*]

[1)] (*School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi* 214122)
[2)] (*Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi* 214122)

Abstract: Aiming at the problem that Siamese tracking algorithm is easy to track drift or loss in complex situations such as target deformation and similar object interference, a target tracking algorithm combining residual connection and channel attention mechanism is proposed. First, the shallow structure features and the deep semantic features extracted from the template branch network are effectively fused through residual connections to improve the model's representational ability. Second, the channel attention module is introduced to make the model adaptively weighted to different semantic target feature channels to improve the generalization ability of the model. Finally, a weight mask based on correlation response values is designed and proposed to increase the weight of similar semantic target loss values during offline training, so that the model is enhanced discrimination of similar semantic targets in end-to-end offline learning. The results from comparative experiments with mainstream tracking algorithms on standard tracking datasets OTB, TempleColor128, VOT2016 and VOT2018 show that the algorithm is highly competitive in tracking accuracy and success rate, with superior real-time performance and reliability.

收稿日期: 2020-04-30; 修回日期: 2020-11-24. **基金项目**: 国家自然科学基金(61806006); 江苏省研究生创新计划(KYLX16_0781); 江苏高校优势学科建设工程资助项目. **邵江南**(1994—), 男, 硕士研究生, CCF 学生会员, 主要研究方向为目标跟踪、深度学习; **葛洪伟**(1967—), 男, 博士, 教授, 博士生导师, 论文通讯作者, 主要研究方向为人工智能与模式识别、机器学习、图像处理与分析.

Key words: object tracking; convolutional neural network; siamese networks; feature fusion; channel attention mechanism

目标跟踪是计算机视觉的重要分支之一, 正随着信息科技的发展在人机交互、智能机器人、自动驾驶、视频监控和智慧城市等领域中得到越来越多的重视和应用. 目标跟踪算法依托于视频帧序列的第 1 帧目标信息, 用于初始化算法模型并完成对后续帧目标的跟踪定位. 尽管视觉跟踪技术在过去数十年中得到了长足的发展, 但由于目标遮挡、尺度变化、外观形变以及相似物体干扰等跟踪环境因素的复杂多变, 且其对跟踪实时性与精度要求高, 仍面临着严峻的挑战.

近年来, 深度学习发展日趋成熟, 在目标跟踪中的应用也越来越广泛. 深度学习能够利用大量已知数据训练网络模型学习对目标特征信息的拟合能力, 能够捕捉目标的深度语义特征, 具有强大的表征能力. 因而众多结合深度卷积神经网络的目标跟踪算法开始涌现, 并吸引大量国内外学者不断研究和探索. MDNet[1]基于 VGG-M 卷积神经网络[2], 采用多域学习的思想, 将每一组跟踪视频序列视为一个独立的域, 并将域相关的判别模块和域无关的特征提取模块分开训练, 提高了跟踪精度的同时降低了过拟合的风险, 但判别模块所使用的全连接层参数量大, 且需要在跟踪进行时定期更新参数, 以维持对视频域目标变化的适应能力, 使得跟踪速度难以应付实时性需求. DeepSRDCF[3]使用 ImageNet-VGG-2048 网络[2]提取的深度特征替代 DCF 算法[4]常用的传统特征, 这种结合了深度学习与相关滤波的跟踪算法, 在一定程度上提高了跟踪精度, 但使用分类数据集预训练的卷积网络难以保留目标的位置和纹理等信息; 这种特征层面的简单替换也导致了深度特征利用不充分和冗余计算, 同样影响了跟踪速度. UDT+[5]则基于 DCFNet[6]提出无监督学习跟踪框架, 通过前向跟踪与反向跟踪结果的一致性损失训练模型, 缓和了特定于跟踪领域的数据集样本数量少的问题, 提高了跟踪精度; 但其忽视背景信息对运动目标追踪的作用, 使得模型难以应付如快速运动和目标遮挡等复杂跟踪环境.

由于基于分类网络的深度跟踪算法存在速度慢且模型提取的特征难以保留位置和纹理信息等问题, 孪生网络(siamese)模型近年来开始被应用于目标跟踪并逐渐占据主导地位: SiamFC[7]是基于全卷积孪生网络的视觉跟踪算法, 其使用 2 个相同结构和参数的网络分支提取目标和候选域特征, 通过卷积操作进行相似性计算, 以估计目标位置, 这种简单而有效的模型结构大大提升了跟踪速度与精度. 此后, 基于 SiamFC 改进的孪生跟踪算法层出不穷: SiamTri[8]将三元损失函数引入 SiamFC, 在迭代训练中挖掘目标与正负样本之间的潜在联系, 从而增强模型的表征能力; SiamRPN[9]将区域建议网络与孪生网络结合, 通过替换传统的尺度金字塔得到更广泛的采样区间, 并将分类分支和回归分支分别用于判别目标和微调模型输出的目标位置; SiamVGG[10]将较深层的 VGG-16 网络[11]用于特征提取, 通过利用更高维度的目标特征获得更鲁棒的特征表达, 从而提高跟踪效果; SiamDW[12]通过设计裁剪残差单元(cropping-inside residual units, CIR)块, 缓和深层网络中所使用的 padding 操作会造成目标空间信息丢失的问题, 从而将 ResNet[13]和 Inception[14]这样更深更宽的网络模型用于目标跟踪中.

这些基于相似性学习的孪生网络模型, 依赖于视频域第 1 帧目标特征的有效性; 而大多数的孪生网络算法都不能充分地利用首帧目标信息, 大多只通过引入更深、更复杂的网络模型提取更高维度的目标特征. 这在一定程度上可以提高跟踪精度, 但也增加了模型训练的复杂度并严重影响了跟踪速度; 且由于卷积特征的平移不变性, 孪生跟踪模型难以应付相似语义物体的干扰.

针对上述问题, 本文提出了一种融合残差连接与通道注意力机制的孪生网络目标跟踪算法, 在 SiamFC 的模板分支上通过残差连接融合浅层结构特征与深层语义特征, 并引入通道注意力模块以充分利用首帧的目标信息; 同时, 设计一种基于相关性响应值的权重掩码, 在离线训练时对难分样本损失值加权, 在增强模型对相似语义物体的辨别力同时, 有效地保持了优越的跟踪精度和实时性.

1　相关工作

1.1　SiamFC 跟踪算法

如图 1 所示, SiamFC 网络模型由 2 个共享权重的分支组成: 模板分支和搜索分支, 分别用于对首帧目标和输入图像提取特征, 并将所提取的特征

输入到互相关层进行相似性计算，以实现对运动目标的定位跟踪.

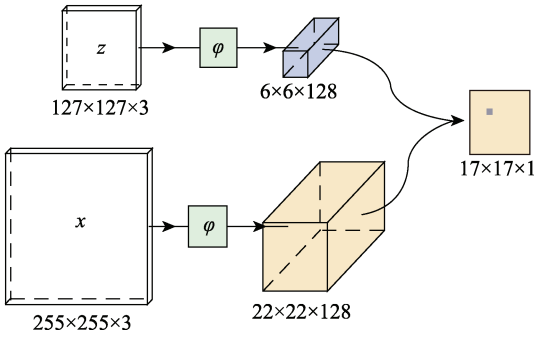


图 1 SiamFC 模型网络结构

SiamFC 算法的关键是离线学习一个相似性度量函数 $f(\cdot)$，用于计算 2 个分支所提取特征的相似度，通过最高响应点预估目标位置，再进行后续操作. 函数 $f(\cdot)$ 如

$$f(z,x)=\varphi(z)*\varphi(x)+b \tag{1}$$

其中，z 为首帧目标图像；x 为输入的搜索图像；$\varphi(\cdot)$ 为各分支对相应图像所提取的深度特征；$*$ 为互相关运算；$b\in \boldsymbol{R}^{n\times n}$ 为各位置点取值的偏置信号. 其中，$\boldsymbol{R}^{n\times n}$ 为 $n\times n$ 的实数矩阵；n 则表示矩阵维度. $f(\cdot)$ 输出 z 和 x 间的相似性响应分值图，该图中最高值点即为目标相对位置.

在离线训练中, SiamFC 通过最小化损失函数

$$L(\boldsymbol{Y}_{n\times n},\boldsymbol{V}_{n\times n})=\sum_i \frac{\ln(1+\exp(-\boldsymbol{Y}_{n\times n}[i]*\boldsymbol{V}_{n\times n}[i]))}{n^2} \tag{2}$$

来获取最优模型参数. 其中，$\boldsymbol{V}_{n\times n}[i]$ 表示模型输出的相似性响应图中的第 i 点的响应值；$Y_{n\times n}[i]\in[0,1]$ 为相应点真实样本类别，其中 1 为正样本中心区域点，其余为 0.

1.2 通道注意力机制

注意力机制广泛应用于目标检测、图像分类和人体姿势估计等多个方面，能够使模型在训练中学习到对空间、特征通道和背景等信息的建模能力，有效地提升卷积神经网络的表征性能. 由于不同特征通道是从不同角度对目标深度信息进行建模，因此针对不同目标各特征通道所发挥的作用不同. 通道间也存在着相互依赖关系，基于此，SENet[15]通过显式地建模特征通道之间的相互依赖关系，并自适应地提取不同通道的权重，显著提升了模型图像分类能力. 在 SENet 的基础上，ECA-Net[16]通过全局平均池化(global average pooling, GAP)和局部化的全连接层(fully connected layers, FC)替代“压缩-激励”操作，同时根据通道间依赖的局部性，把单一通道的依赖关系提取依据特征维度限定在相邻的 k (k <9)个通道以内，显著地提升了模型对通道信息的建模速度.

2 本文算法

本文算法在 SiamFC 基础上通过残差连接融合模板分支网络不同层所提取的深度特征，并结合 ECA-Net 引入快速通道注意力机制(effcient channel attention, ECA)，增大对首帧目标信息的利用率，同时设计了基于相关性响应的权重掩码，以增加模型对相似物体的辨别力. 图 2 所示为本文算法的模型网络框架图.

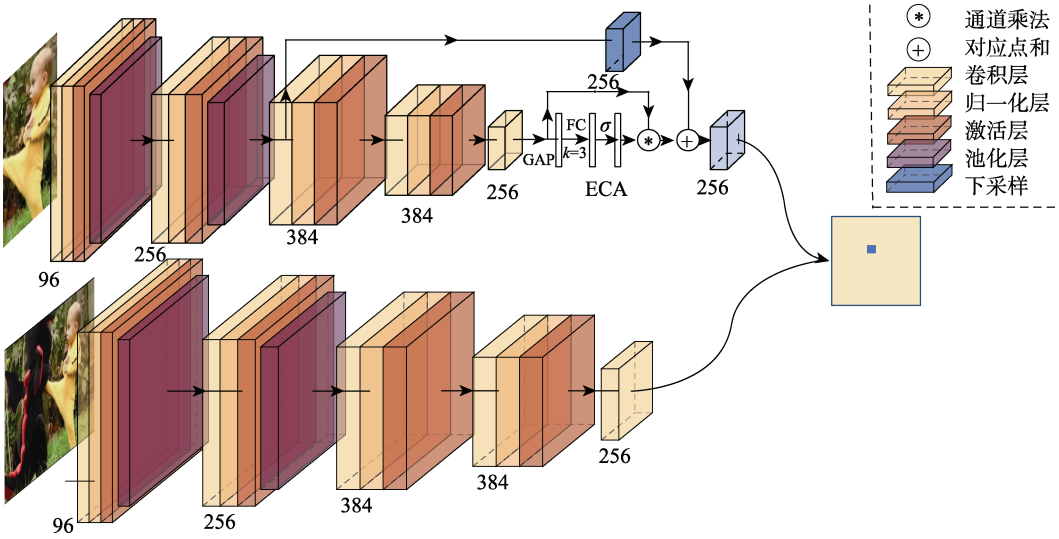


图 2 本文算法模型网络框架

2.1　残差连接

卷积神经网络通过逐层计算并前向传播最终生成目标高维深度特征. 如图 3 所示，不同卷积层对表征目标信息的侧重点不同，深层网络有利于提取目标语义特征，以实现更高置信分类；浅层更有利于保留目标的位置、轮廓、尺度和颜色等结构特征. 但大多数的深度模型只利用最终层的输出特征表征样本，这在一定程度上造成了模型性能的损失和浪费.

由于目标跟踪的属性与分类任务存在本质区别，需要更深层的语义信息进行候选样本判别的同时，也需要更丰富的浅层结构特征，以实现更精确的目标定位. 因此，轻量级卷积神经网络 AlexNet[17]难以应付复杂的跟踪环境，而如 VGG 和 ResNet 等结构更深的卷积网络大大影响了跟踪速度. 基于此，本文使用改进后的 AlexNet 作为模型骨干网络，通过融合不同层卷积特征，在最大限度地维持实时性能的同时，增大模型的表征能力.

根据特征维度的一致性，本文算法使用最近邻点插值法对第 2 层卷积特征 $\boldsymbol{F}_{2,256\times12\times12}$ 进行下采样操作，通过残差连接将得到的采样后的特征 $\boldsymbol{F}_{2',256\times6\times6}$ 与最后一层卷积特征 $\boldsymbol{F}_{5,256\times6\times6}$ 进行线性融合. 通过端到端的离线训练，模型能够在一次次迭代中学习结构特征和语义特征在融合中的相应权重，并在提取语义信息的同时，保留目标的结构信息，通过结合目标的语义与结构特征进行相似性判别，从而实现鲁棒目标跟踪.

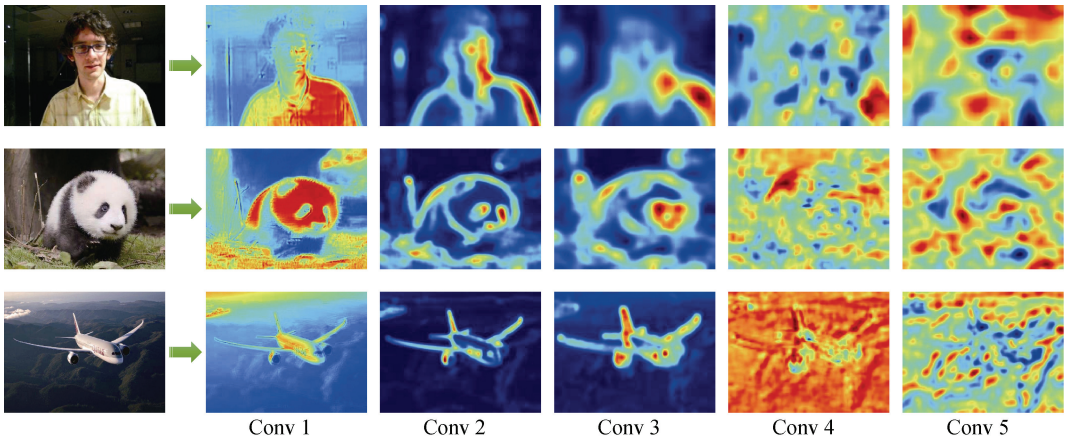


图 3　各卷积层输出特征可视化图

2.2　通道注意力模块

不同卷积核提取的不同特征通道，对应不同类型的视觉模式和语义属性. 对不同目标，有些特征通道可能更为重要，而有些通道可能完全无关；这种无关于目标语义的特征通道常常会影响模型的相似性计算过程，进而影响跟踪结果.

如图 2 所示，本文在对首帧目标的处理上，结合 ECA 设计通道注意力模块，对模板分支所提取到的首帧目标特征，通过 GAP 和 FC 确定相应特征通道权重，并捕捉各特征通道和其相邻 k=3 个通道之间的依赖关系. 通过这种通道注意力机制的引入，模型能够在端到端训练中学习对不同语义目标的不同特征通道的重要性解读能力，从而自适应地对特征通道加权，以充分挖掘并利用首帧目标语义信息，提高模型的表征能力.

视频序列的第 1 帧目标信息对目标跟踪至关重要，除了外观和位置等结构数据，还能提供整个视频域的目标语义信息，这种语义信息能够被深度卷积神经网络所提取. 而大多数的孪生跟踪模型都不能充分地利用首帧目标信息，只用于求取其在随后帧中的相似性响应. 基于此，本文算法所使用的残差连接和注意力模块均只应用于模板分支，即对首帧目标的特征提取阶段，避免在双分支网络中因简单地增加网络层而影响跟踪速度. 得益于这种非对称结构的网络模型设计，本文算法能够在显著地降低过拟合风险的同时，提取更高维度(256×6×6，256×22×22)的特征，以产生性能增益，在增大了对首帧目标信息的利用的同时，最大程度地保障了跟踪实时性.

2.3　损失掩码

如式(2)所示，SiamFC 在训练时直接通过响应值图与目标分布图计算逻辑损失值，这样简单的损失函数虽能达到一定的效果，但是将所有的负样本都不加区分地认为是普通负样本，不利于模型学习区分相似语义或外观属性的干扰物体，即难分负样本的能力.

为了让模型在离线训练时增强对这种相似目标的区分能力, 本文首次设计并实现了一种基于相似性响应图的损失掩码(loss-mask), 旨在每一次迭代损失值计算时, 通过提高难分样本损失值的权重, 使模型将参数优化方向部分转移到区分难分样本中来. 掩码 $\boldsymbol{M}_{n\times n}$ 的计算方式为

$$\begin{cases}\boldsymbol{M}_{n\times n}=\boldsymbol{N}(\boldsymbol{R}(\boldsymbol{V}_{n\times n}-\boldsymbol{V}_{n\times n}[t]))\\ \boldsymbol{N}(\boldsymbol{S})=-\dfrac{\boldsymbol{S}-\min(\boldsymbol{S})}{\max(\boldsymbol{S})-\min(\boldsymbol{S})+e^{-8}}\end{cases}\tag{3}$$

其中, $\boldsymbol{V}_{n\times n}$ 为模型输出的相似性响应值图, 大小为 $n\times n$; $\boldsymbol{V}_{n\times n}[t]$ 为模型对真实目标点的响应值; $\boldsymbol{R}(\cdot)$ 为激活函数, 用于筛选难分样本, 只保留响应值大于真实目标点的候选; $\boldsymbol{N}(\cdot)$ 为归一化函数, 通过对式中 $\boldsymbol{S}=\boldsymbol{R}(\boldsymbol{V}_{n\times n}-\boldsymbol{V}_{n\times n}[t])$ 的归一化操作, 能有效地避免局部点损失权重过大而掩盖其余位置对训练过程的影响.

$\boldsymbol{M}_{n\times n}$ 能够在提高跟踪模型性能的同时, 不会对在线跟踪速度带来任何损失, 且由于计算复杂度低、相关参数少, 对离线训练的影响同样较小. 改进后的损失函数为

$$L(\boldsymbol{Y}_{n\times n},\boldsymbol{V}_{n\times n})=\sum_i\left((\boldsymbol{M}_{n\times n}[i]\times\mu)\times\frac{\ln(1+\exp(-\boldsymbol{Y}_{n\times n}[i]*\boldsymbol{V}_{n\times n}[i]))}{n^2}\right)\tag{4}$$

其中, μ 为超参数, 用于控制掩码在损失计算中的影响系数; $\boldsymbol{M}_{n\times n}[i]$ 为损失掩码 $\boldsymbol{M}_{n\times n}$ 中第 i 点的对应值.

2.4 算法流程

在线跟踪时, 本文算法类似于 SiamFC 等常规孪生网络模型, 不在线调整模型参数, 也无需保存历史帧的目标特征, 通过利用跟踪任务中目标变化的连续性和位置变化的局部性, 实现准确而高效的视频目标定位. 算法的主要步骤如下:

Step1. 输入视频帧序列和第 1 帧图像 N_1 的目标位置(X_1, Y_1, H_1, W_1).

Step2. 通过模板分支提取 N_1 的目标特征 F_1.

Step3. 对于第 t 帧图像 N_t, 截取(X_{t-1}, Y_{t-1}, $3H_{t-1}+W_{t-1}$, $3W_{t-1}+H_{t-1}$)区域作为搜索域, 将搜索域分别放缩尺度 S_1, S_2, S_3, 并双三次插值为 255×255 大小, 其中(S_1, S_2, S_3)=$1.0572^{(-1,0,1)}$.

Step4. 使用搜索分支提取 Step3 中所得的搜索域特征, 并分别计算与 F_1 的相似度响应, 得到响应图 I_{Response_1}, I_{Response_2}, I_{Response_3}.

Step5. 计算最大响应值所对应的响应图 I_{Response_k} 和放缩尺度 S_k ($k\in\{1, 2, 3\}$).

Step6. 利用余弦窗对 I_{Response_k} 进行较大位移响应值抑制.

Step7. 通过 S_k 和 Step6 处理后的 I_{Response_k} 的最大值位置计算当前帧目标位置(X_t, Y_t, H_t, W_t).

Step8. 重复 Step3~Step7 直至所有帧跟踪结束.

3 实验结果与分析

本文算法基于 CUDA 10.0 深度学习框架和 PyTorch 1.2.0 编程语言实现, 实验操作系统为 Ubuntu 16.04, 内存 64 GB, CPU 为 Intel i9-9900x 3.5 GHz, GPU 为 NVIDIA RTX2080Ti.

为验证模型的有效性、实时性以及泛化能力, 分别在 OTB2013[18], OTB50, OTB100[19], TempleColor128[20], VOT2016[21]和 VOT2018[22]数据集上与主流跟踪算法进行了对比实验, 统计各数据集上所有帧的平均跟踪结果作为模型性能衡量依据.

3.1 实验设置

模型离线训练于 GOT-10K 标记数据集[23], 学习率初始值为 0.01, 衰减系数为 0.868 5; 训练 50 个 epoch; 模型激活函数为 Mish[24]; 设置式(4)中的 μ=3, 余弦窗权重系数设置为 0.235 6.

为保证实验的公平性, 实验中所用对比跟踪算法的实验结果均来源于相关作者论文所给或使用其开源代码和参数实际运行所得.

3.2 评价指标

本文主要采用跟踪精度(precision, Prec)和跟踪成功率作为评价指标对比各跟踪算法综合性能.

(1) 跟踪成功率是通过计算目标预测框 R_{t} 和真实边界框 R_{a} 的重叠率(intersection over union, IoU)曲线下面积(area under curve, AUC)所得. IoU 计算公式为

$$\text{IoU}=\frac{|R_{\text{t}}\cap R_{\text{a}}|}{|R_{\text{t}}\cup R_{\text{a}}|}\tag{5}$$

其值越大, 表示跟踪算法的成功率越高. 通常, 当 IoU>0.5 时, 可视为该帧目标被成功定位.

(2) 跟踪精度则可通过计算 R_{t} 与 R_{a} 的中心位置(x_{t}, y_{t}), (x_{a}, y_{a})间的欧几里得度量

$$\varepsilon=\sqrt{(x_{\text{t}}-x_{\text{a}})^2+(y_{\text{t}}-y_{\text{a}})^2}\tag{6}$$

来衡量. ε 越小, 表示跟踪精度越高. 取 ε 在 20 个像素点以内的帧的百分比作为模型综合跟踪精度.

3.3　定量分析

3.3.1　在 OTB 数据集上测试

OTB2013，OTB50，OTB100 分别含 51，50 和 100 个视频跟踪序列，是目标跟踪领域常用的标准评估测试集．本文基于上述数据集分别与主流跟踪算法 UDT+[5]，SiamRPN[9]，ACFN[25]，SiamTri[8]，DCFNet[6]，SRDCF[26]，Staple[27]和 SiamFC[7]等进行了对比实验，实验结果如图 4 所示．

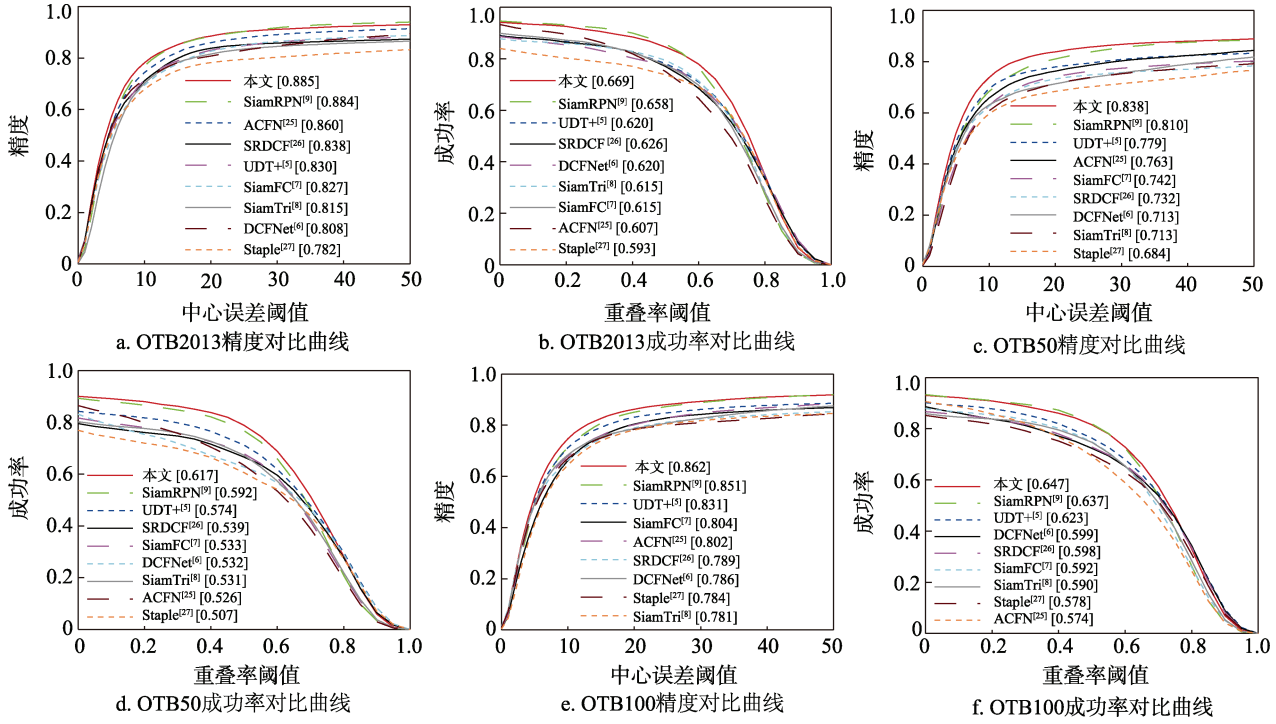


图 4　9 种算法在 3 个数据集上的跟踪结果评估曲线

从图 4 可以看出，本文算法在所有 OTB 标准数据集上均较对比算法展现出了更优的跟踪精度和成功率．在视频序列最多的 OTB100 中，本文算法综合 Prec=86.2%，AUC=64.7%，分别较 SiamFC 提高 5.8%和 5.5%，较其余的最优对比算法提高 1.1%和 1.0%.

针对 OTB 数据集所包含的 11 个不同的视频属性：光照变化(illumination variation, IV)、尺度变化(scale variation, SV)、目标被遮挡(occlusion, OCC)、目标形变(deformation, DEF)、运动模糊(motion blur, MB)、快速运动(fast motion, FM)、平面旋转(in-plane rotation, IPR)、平面外旋转(out-of-plane rotation, OPR)、目标出视野(out of view, OV)、低分辨率(low resolution, LR)和背景相似物干扰(background clutter, BC)，表 1 定量展示了本文算法和各对比跟踪算法在应对这些复杂跟踪因素下的平均 Prec.

表 1　9 种算法在 OTB100 数据集上对 11 个视频属性的 Prec 定量对比

属性	视频数	UDT+[5]	DCFNet[6]	ACFN[25]	Staple[27]	SRDCF[26]	SiamTri[8]	SiamRPN[9]	SiamFC[7]	本文
OPR	63	0.803	0.762	0.777	0.738	0.742	0.763	*0.855*	0.797	**0.861**
IV	37	0.775	0.743	0.784	0.778	0.786	0.752	**0.873**	0.759	*0.851*
SV	63	0.802	0.753	0.761	0.724	0.741	0.752	*0.846*	0.775	**0.849**
OCC	48	0.787	0.770	0.751	0.724	0.730	0.730	*0.791*	0.740	**0.840**
DEF	43	0.798	0.722	0.768	0.747	0.728	0.683	*0.837*	0.785	**0.841**
MB	29	0.813	0.712	0.730	0.699	0.767	0.727	**0.821**	0.772	*0.817*
FM	39	0.779	0.740	0.757	0.710	0.769	0.763	*0.793*	0.779	**0.807**
IPR	51	0.772	0.779	0.784	0.768	0.745	0.774	**0.859**	0.807	*0.851*
OV	14	**0.780**	0.697	0.690	0.668	0.597	0.723	0.728	0.697	*0.731*
LR	9	0.788	0.713	0.623	0.610	0.655	*0.897*	0.870	0.794	**0.914**
BC	31	*0.813*	0.734	0.769	0.749	0.775	0.715	0.803	0.710	**0.824**
OTB100	100	0.831	0.786	0.802	0.784	0.789	0.781	*0.851*	0.804	**0.862**

注．加粗字体为每行最优值，斜体为每行次优值．

从表 1 可以看出，在 OTB 数据集的 11 个复杂跟踪因素中，本文算法对其中 7 个跟踪因素保持了最优性能，其余则均取得了次优性能. 在应对上述所有跟踪因素时均大幅领先于 SiamFC，其中，在 OCC, DEF 和 BC 情况下, Prec 分别较 SiamFC 提升 10%, 5.6%和 11.4%.

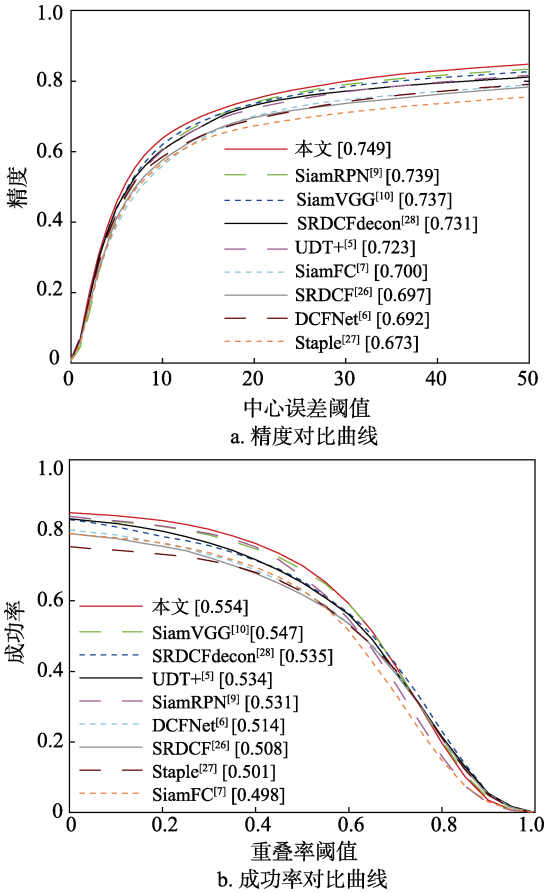
3.3.2 在 TempleColor128 数据集上测试

TempleColor128 包含 128 个彩色视频序列，更加贴近于现实场景的跟踪环境，其视频复杂跟踪因素和模型评估指标均与 OTB 数据集一致. 为了进一步验证模型的泛化能力，本文算法与 UDT+[5], DCFNet[6], SiamFC[7], SiamRPN[9], SiamVGG[10], SRDCF[26], Staple[27]和 SRDCFdecon[28]在 TempleColor128 进行了对比实验，实验结果如图 5 所示.

从图 5 可以看出，本文算法在视频帧序列数量更多的 TempleColor128 上仍保持了优越的跟踪性能，在综合 Prec 和 AUC 上分别较最优对比模型 SiamRPN, SiamVGG 提升 1.0%和 0.7%，较 SiamFC 分别提升了 4.9%和 5.6%.

3.3.3 在 VOT2016 和 VOT2018 数据集上测试

VOT2016 和 VOT2018 分别含有 60 组高分辨率视频帧序列，采用可旋转的边界框更精确地标注目标位置，且使用了不同于 OTB 和 TempleColor128 的评价指标：准确率 A、鲁棒性 R 和期望平均重叠率(expected average overlap, EAO)，分别衡量跟踪算法的跟踪精度、失败率和综合跟踪性能，对模型的要求更为严格，性能评估也更为全面. 本文基于上述数据集与 SiamRPN[9], SiamDW[12], GradNet[29], SA-Siam[30], MemTrack[31], Staple[27], UDT+[5], SRDCF[26], TADT[32], SiamVGG, C-RPN[33] 和 SiamFC[7]进行了对比实验，实验结果如表 2 所示.



a. 精度对比曲线

b. 成功率对比曲线

图 5 不同算法在 TempleColor128 上的跟踪结果评估曲线

表 2 13 种算法在 2 个数据集上的实验结果

算法	VOT2016			VOT2018		
	$A\uparrow$	$R\downarrow$	EAO↑	$A\uparrow$	$R\downarrow$	EAO↑
SiamRPN[9]	*0.56*	**0.26**	0.344	0.49	0.46	0.244
TADT[32]	0.55	0.33	0.299	0.51	0.42	0.235
GradNet[29]				0.51	0.38	0.247
MemTrack[31]	0.53	0.37	0.273	0.52	*0.36*	0.248
Staple[27]	0.54	0.38	0.295	0.52	0.69	0.169
UDT+[5]	0.53	0.31	0.301			
SRDCF[26]	0.54	0.42	0.247	0.49	0.97	0.119
SA-Siam[30]	0.54		0.291	0.50	0.46	0.236
SiamVGG[10]	*0.56*		0.351	0.52		0.286
C-RPN[33]			**0.363**			*0.289*
SiamDW[12]	0.54	0.30	0.303	**0.54**	0.40	0.270
SiamFC[7]	0.53	0.46	0.235	0.51	0.48	0.234
本文	**0.58**	*0.27*	*0.362*	*0.53*	**0.34**	**0.299**

注. 加粗字体为每列最优值，斜体为每列次优值.

从表 2 可以看出，本文算法在 VOT2016 和 VOT2018 上均展现出了良好的跟踪性能. 其中，在 VOT2016 上本文算法的 A 值优于所有的对比算法，虽然 R 值稍落后于 SiamRPN，但其余评估指标均明显优于后者；虽然 EAO 值较 C-RPN 低 0.1%，但在难度更大的 VOT2018 上本文算法提升了 1.0%，较 SiamFC 在 VOT2016 的 A, R 和 EAO 上分别提升 5.0%，19.0%和 12.7%；在 VOT2018 上，本文算法虽然在 A 上落后于 SiamDW 1.0%，但在 R 和 EAO 上分别较后者提高 6.0%和 2.9%，且较其他模型则在 A, R 和 EAO 上均保持了领先，较 SiamFC 在上述评价指标上分别提升 2.0%, 14.0%和 6.5%.

从定量分析对比实验可以看出，本文算法有着优越的跟踪性能和泛化能力，并在应对多种复杂跟踪因素时保持了良好的稳定性和可靠性.

3.4 定性分析

为了更直观地对比本文算法和各种对比跟踪算法在应对复杂跟踪环境时的实际跟踪性能，选取

OTB100 数据集上较典型的 4 组视频序列(Human3, MotorRolling, Skating 和 Soccer)与 OTB 上综合表现最好的 SiamRPN[9], UDT+[5], SiamFC[7], ACFN[25]和 SRDCF[26]进行了定性对比实验，图 6 着重展示了各跟踪算法在应对遮挡、相似物体干扰、目标形变等因素下的实际表现.

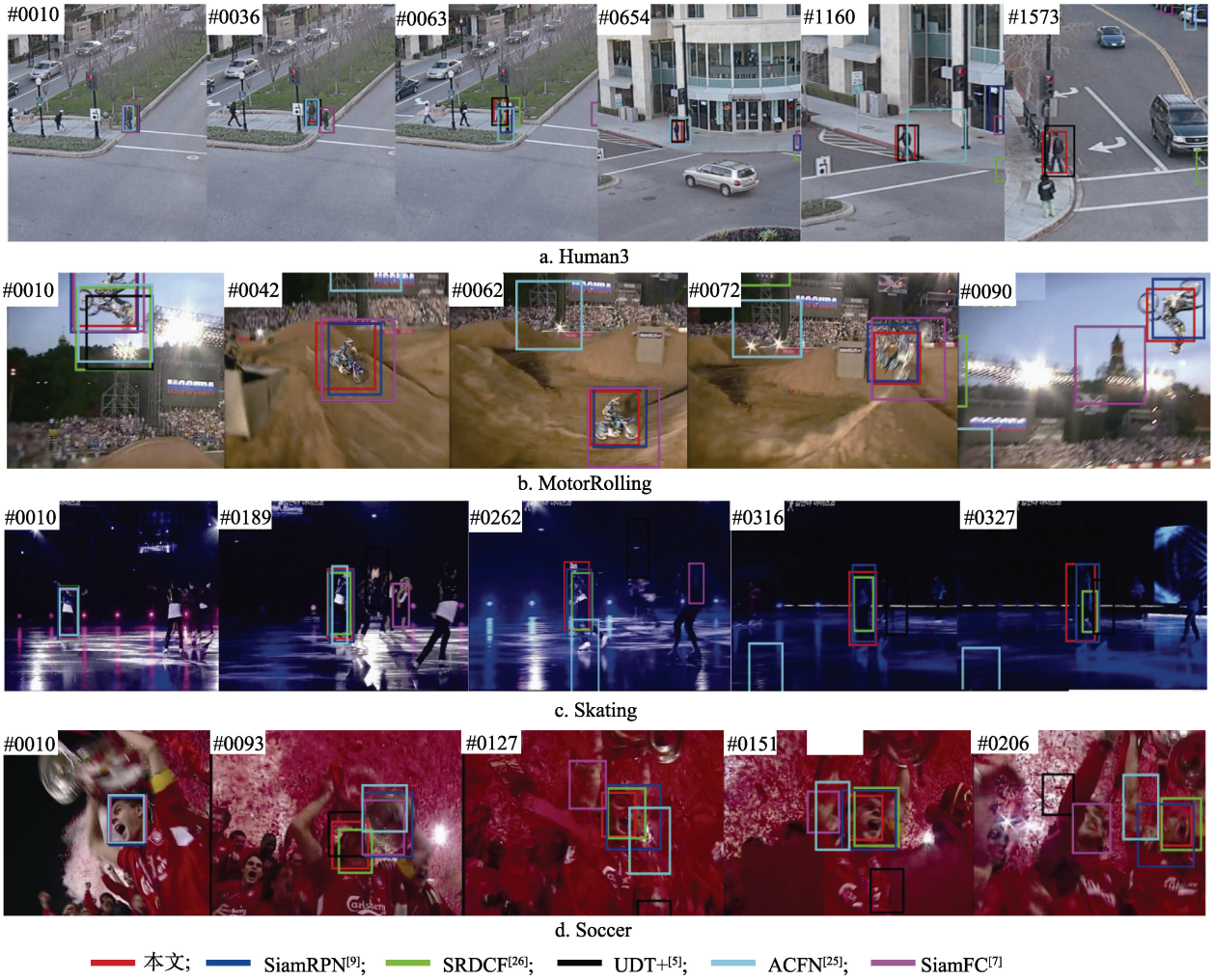


图 6　6 种算法在 4 组视频序列下的定性对比结果

图 6a 所示为对行人这种小目标进行跟踪，在第 36 帧，目标行人与他人交汇，SiamFC 随即产生跟踪漂移并持续失跟至视频帧结束；在第 63 帧和第 654 帧，行人被交通灯遮挡后重现，此时只有本文算法和 UDT+能第一时间定位目标位置；在第 1 160 帧和之后的帧序列里，只有本文算法能精确地锁定目标位置和轮廓信息.

图 6b 所示为摩托车在第 10 帧做出空翻动作，此时目标高速位移和形变，UDT+[5]，ACFN[25]和 SRDCF[26]均在此后失去了跟踪能力；在第 42 帧，第 62 帧和第 72 帧序列，车高速俯冲，此时本文算法能够比 SiamRPN 和 SiamFC 更准确地定位目标位置和轮廓；SiamFC 在第 90 帧车身再次翻转时跟踪失败.

图 6c 所示为在昏暗场景下对滑冰者进行位置跟踪，此时视频分辨率低，场地相似目标较多，对跟踪定位有很大挑战. 在第 189 帧，目标靠近其他滑冰者，此时 SiamFC，UDT+均被相似目标干扰；在第 262 帧和第 316 帧，目标高速旋转，背景灯光也随之变化，此时 ACFN 彻底丢失目标位置信息；在第 327 帧 SRDCF 也同样失去了对目标的横纵比和轮廓的建模能力，而本文算法和 SiamRPN 始终保持了对目标的定位跟踪，且本文算法较后者有着更准确的轮廓信息表达.

图 6d 所示为穿着红色衣服的运动员在第 93 帧昂头举起奖杯，此时 SiamRPN, SiamFC 和 ACFN 均由于目标变模糊和背景干扰产生跟踪漂移；在第 127 帧和第 151 帧，目标被部分遮挡，UDT+随之

跟踪失败; 在第 206 帧, 左侧出现相似目标, SiamFC 完全被干扰, 而本文算法能够始终精确地锁定目标位置.

从上述的分析可见, 本文算法能够在应对目标遮挡、外观形变和相似目标干扰等复杂跟踪环境时准确定位目标位置, 并对目标横纵比和轮廓信息有着更精确的建模能力.

3.5 实时性分析

为进一步对比分析本文算法模型的实时性能, 基于 OTB100 数据集在本实验环境下与前沿孪生网络衍生模型 SiamVGG[10], SiamDW[12], SiamTri[8], RASNet[34], TADT[32], SiamRPN[9]和 SiamFC[7]进行 AUC 和跟踪帧率的对比实验, 结果如表 3 所示.

表 3 8 种算法在 OTB100 数据集上的对比

算法	AUC	帧率/(帧·s^{-1})
RASNet[34]	0.642	83
TADT[32]	**0.656**	92
SiamDW[12]	*0.654*	97
SiamTri[8]	0.590	**164**
SiamVGG[10]	*0.654*	59
SiamRPN[9]	0.637	135
SiamFC[7]	0.592	**164**
本文	0.647	*154*

注. 加粗字体为每列最优值, 斜体为每列次优值.

从表 3 可以看出, 本文算法的 AUC 和帧率均优于 RASNet, SiamRPN; 在帧率略低于 SiamFC 和 SiamTri 的情况下, AUC 值大幅提升; 尽管 AUC 较 SiamDW, SiamVGG 和 TADT 分别落后 0.7%, 0.7% 和 0.9%, 但本文算法的帧率大幅领先后者. 可见, 本文算法能够更好地均衡目标定位能力与跟踪速度, 并在多种应用环境下满足实时性需求, 具有良好的综合跟踪性能.

3.6 消融实验

为验证本文基于 SiamFC 的 3 个主要改进点: 残差连接(RC)、通道注意力机制(CA)和损失值权重掩码(LM)的有效性, 在 OTB2013 上对所提模型进行了消融实验, 各模块 AUC 和 Prec 数值对比如表 4 所示.

表 4 本文算法在 OTB2013 数据集消融实验结果

算法	AUC	Prec
SiamFC	0.615	0.827
SiamFC+LM	0.626	0.858
SiamFC+RC	0.652	0.858
SiamFC+RC+CA	0.661	0.877
SiamFC+RC+CA+LM	0.669	0.885

实验结果表明, 残差连接、通道注意力机制和损失值权重掩码均能有效地提高跟踪算法的模型性能, 且三者能够协同提高跟踪精度和成功率.

4 结 语

本文提出了一种融合残差连接和通道注意力机制的 Siamese 目标跟踪算法, 通过在模型训练时提高难分样本损失值的权重, 以及在提取高维语义特征时保留结构信息和自适应改变目标特征的通道权重, 提升了模型对相似属性目标的区分能力和在复杂跟踪场景下的泛化能力. 实验结果验证了本文算法在跟踪速度、精度和成功率上的优越性能, 以及在应对目标形变和相似物体干扰等因素时的良好鲁棒性. 下一步的研究将探索在最大程度保留跟踪实时性的同时, 挖掘并利用历史帧时空域信息, 提高模型在长时目标跟踪时的定位精度和稳定性.

参考文献(References):

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

[2] Chatfield K, Simonyan K, Vedaldi A, *et al*. Return of the devil in the details: Delving deep into convolutional nets[OL]. [2020-04-30]. https://arxiv.org/abs/1405.3531

[3] Danelljan M, Häger G, Khan F S, *et al*. Convolutional features for correlation filter based visual tracking[C] //Proceedings of the IEEE International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2015: 621-629

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

[5] Wang N, Song Y B, Ma C, *et al*. Unsupervised deep tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1308-1317

[6] Wang Q, Gao J, Xing J L, *et al*. DCFNet: discriminant correlation filters network for visual tracking[OL]. [2020-04-30]. https://arxiv.org/abs/1704.04057

[7] Bertinetto L, Valmadre J, Henriques J F, *et al*. Fully-convolutional Siamese networks for object tracking[C] // Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2016: 850-865

[8] Dong X P, Shen J B. Triplet loss in Siamese network for object tracking[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 472-488

[9] Li B, Yan J 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. Los Alamitos: IEEE Computer Society Press, 2018: 8971-8980

[10] Li Y H, Zhang X F. SiamVGG: visual tracking using deeper Siamese networks[OL]. [2020-04-30]. https://arxiv.org/abs/1902.02804

[11] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[OL]. [2020-04-30]. https://arxiv.org/abs/1409.1556

[12] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 4586-4595

[13] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[14] Szegedy C, Liu W, Jia Y Q, *et al*. Going deeper with convolutions[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 1-9

[15] Hu J, Shen L, Albanie S *et al*. Squeeze-and-excitation networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 2011-2023

[16] Wang Q L, Wu B G, Zhu P F, *et al*. ECA-Net: efficient channel attention for deep convolutional neural networks[OL]. [2020-04-30]. https://arxiv.org/abs/1910.03151

[17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90

[18] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2013: 2411-2418

[19] Wu Y, Lim J, Yang M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834-1848

[20] Liang P P, Blasch E, Ling H B. Encoding color information for visual tracking: Algorithms and benchmark[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5630-5644

[21] Kristan M, Leonardis A, Matas J, *et al*. The visual object tracking VOT2016 challenge results[C] //Proceedings of European Conference on Computer Vision Workshops. Heidelberg: Springer, 2016: 777-823

[22] Kristan M, Leonardis A, Matas J, *et al*. The sixth visual object tracking VOT2018 challenge results[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 3-53

[23] Huang L H, Zhao X, Huang K Q. GOT-10K: a large high-diversity benchmark for generic object tracking in the wild[OL]. [2020-04-30]. https://ieeexplore.ieee.org/document/8922619/authors#authors

[24] Misra D. Mish: a self regularized non-monotonic neural activation function[OL]. [2020-04-30]. https://arxiv.org/abs/1908.08681

[25] Choi J, Chang H J, Yun S, *et al*. Attentional correlation filter network for adaptive visual tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 4828-4837

[26] Danelljan M, Häger G, Khan F S, *et al*. Learning spatially regularized correlation filters for visual tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2015: 4310-4318

[27] Bertinetto L, Valmadre J, Golodetz S, *et al*. Staple: complementary learners for real-time tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1401-1409

[28] Danelljan M, Häger G, Khan F S, *et al*. Adaptive decontamination of the training set: a unified formulation for discriminative visual tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 1430-1438

[29] Li P X, Chen B Y, Ouyang W L, *et al*. GradNet: gradient-guided network for visual object tracking[C] //Proceedings of the IEEE International Conference on Computer Vision Workshops. Los Alamitos: IEEE Computer Society Press, 2019: 6161-6170

[30] He A F, Luo C, Tian X M, *et al*. A twofold Siamese network for real-time object tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 4834-4843

[31] Yang T Y, Chan A B. Learning dynamic memory networks for object tracking[C] //Proceedings of European Conference on Computer Vision. Heidelberg: Springer, 2018: 153-169

[32] Li X, Ma C, Wu B Y, *et al*. Target-aware deep tracking[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 1369-1378

[33] Fan H, Ling H B. Siamese cascaded region proposal networks for real-time visual tracking[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2019: 7944-7953

[34] Wang Q, Teng Z, Xing J L, *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. Los Alamitos: IEEE Computer Society Press, 2018: 4854-4863