文章编号：1005-5630(2021)01-0014-07 **DOI**: 10.3969/j.issn.1005-5630.2021.01.003

基于区域提案孪生网络的优化目标跟踪算法

秦晓飞[1]，张一鹏[2]，陈浩胜[1]，李 夏[1]，何致远[1]
（1. 上海理工大学 光电信息与计算机工程学院，上海 200093；
2. 上海理工大学 机械工程学院，上海 200093）

摘要：为了更好地对目标的尺度进行实时估计，避免多尺度测试及提高目标跟踪的速度和精度，提出了一种新的优化目标跟踪算法。通过将跟踪效果比较好的区域提案网络引入普通的孪生网络，并在算法中引进条形池化模块和高效通道注意力模块，应对物体的尺度差异和跟踪过程中较为剧烈的形变。提出的算法在 OTB100 数据集上取得了 0.833 的准确度和 0.658 的成功率，在 VOT2016 数据集上取得了 0.411 的 EAO 指数，在 VOT2019 数据集上取得了 0.275 的 EAO 指数。

关键词：孪生网络；区域提案网络；条形池化；通道注意力
中图分类号：TP 391 **文献标志码**：A

Optimization of target tracking algorithm based on region proposal Siamese network

QIN Xiaofei[1]，ZHANG Yipeng[2]，CHEN Haosheng[1]，LI Xia[1]，HE Zhiyuan[1]
（1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China;
2. School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: In order to estimate the target scale in real time, avoid multi-scale test and improve the speed and accuracy of target tracking, a new optimized target tracking algorithm is proposed. By introducing the regional proposal network with good tracking effect into the common siamese network, and introducing the strip pooling module and the efficient channel attention module in the algorithm, we can deal with the scale difference of objects and the severe deformation in the tracking process. The proposed algorithm achieves 0.833 accuracy and 0.658 success rate on OTB100 dataset, 0.411 EAO index on VOT2016 dataset, and 0.275 EAO index on VOT2019 dataset.

Keywords: siamese network；regional proposal network；strip pooling；channel attention

收稿日期：2020-07-22
基金项目：上海市人工智能计划（2019RGZN01077）
作者简介：秦晓飞（1982—），男，高级工程师，研究方向为人工智能算法。E-mail：xiaofei.qin@foxmail.com

引　言

目标跟踪是计算机视觉中的一个重要分支，在人机交互、自动驾驶等领域都有着广泛的应用。目标跟踪的一般要求是仅仅根据第一帧中给出的边界框，就能准确地估计目标在后续帧中的位置和尺度。由照明、变形、遮挡、旋转和运动模糊引起的外观差异都是很大的挑战，而且跟踪速度在实际应用中也是必须考虑的一个方面。通常，实时跟踪的帧率至少为 25 帧/s。

近几年，目标跟踪技术迅速发展，涌现出了大批优秀的跟踪算法。在普通的卷积神经网络中，卷积核或者池化核是正方形的，这对于长宽比比较接近的目标进行采样时，可以取得较好的效果。然而，当物体的长宽比比较悬殊的时候，网络往往显得比较乏力，特别是在骨干网络中，这样的操作直接影响了后续的信息处理。本文通过引入条形池化模块[1]来增加网络对于细长物体的采样能力，同时，由于是轻量级模块，计算量和参数量的增加微乎其微。

区域提案孪生网络[2]的分类分支作用是将目标的前景区域和背景区域分开，从而获取跟踪物的位置信息，给边界框回归提供参考，从而得到高质量的预测框。所以分类分支的分类性能，直接影响了整个跟踪器的跟踪效果，若能抑制干扰信息，就可得到一个更加具有判别力的分类器。通道注意力模块[3-7]的作用是根据神经网络的不同输入对不同部分适应性地分配权重。本文引进了高效通道注意力模块[8]，可以从神经网络的通道维度有效抑制干扰信息，而使有用的信息得到有效保留。同时还在 OTB100[9]、VOT2016[10]和 VOT2019[11]等数据集中对提出的方法进行了评估。

1　目标尺度感知的区域提案孪生网络目标跟踪

1.1　网络整体框架

网络整体框架如图 1 所示，由 2 个子网络构成，分别是左侧的孪生子网络和右侧的提案生成子网络。视频的第一帧从模板分支输入，而后续帧从搜索分支输入，2 个子窗口的片段经相同的网络（网络结构和参数均相同）AlexNet[12]后，再将模板分支的特征图送入条形池化模块（即 SPM）[1]进行进一步的处理。

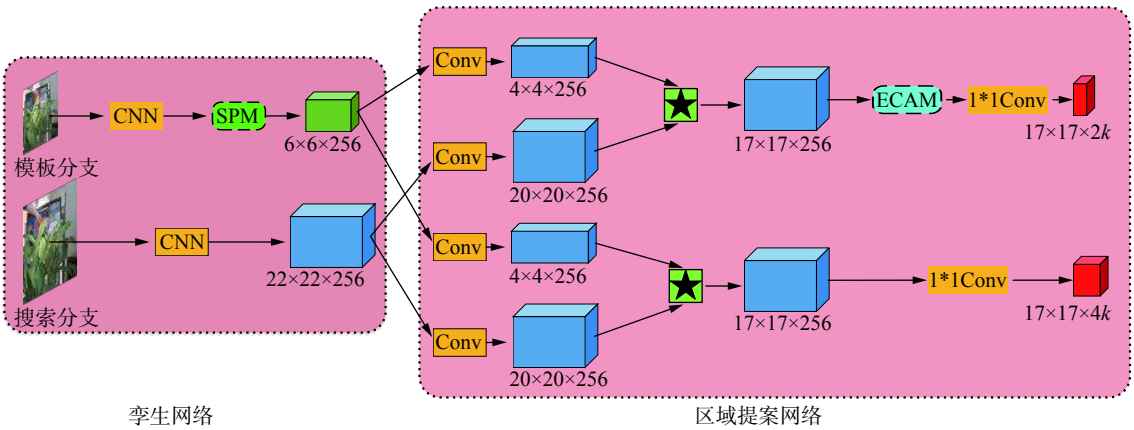


图 1　本文网络的整体框架

Fig. 1　The overall framework of this paper

提案生成子网络由分类分支和回归分支 2 个部分构成：前者负责前景和背景分类，在每个锚点的位置生成 $2k$ 个得分，分别是 k 个前景得分和 k 个背景得分，用于后续的锚点框筛选；后者负责生成锚点框，在每个锚点的中心位置预设了 5 个尺度的锚点框，长宽比分别为 1∶1、1∶2、2∶1、1∶3、3∶1。网络的输出是每个框的边界框的中心点的横坐标、纵坐标和宽度、高度的修正量，而每个锚点框与分类分支的得分一一对应，这样每个框就有了得分，然后根据惩罚函数得到最终的预测框。

1.2　条形池化模块

在目标跟踪任务中，目标物的尺度是实时变化且未知的，很有可能会出现物体的边界框长宽比悬殊的情况。这个时候，普通的卷积神经网络就不能很好地采样，会严重影响跟踪算法的尺度估计。本文引入的条形池化模块[1]，如图 2 所示，通过在狭长区域进行采样，很好地缓解了这个问题。

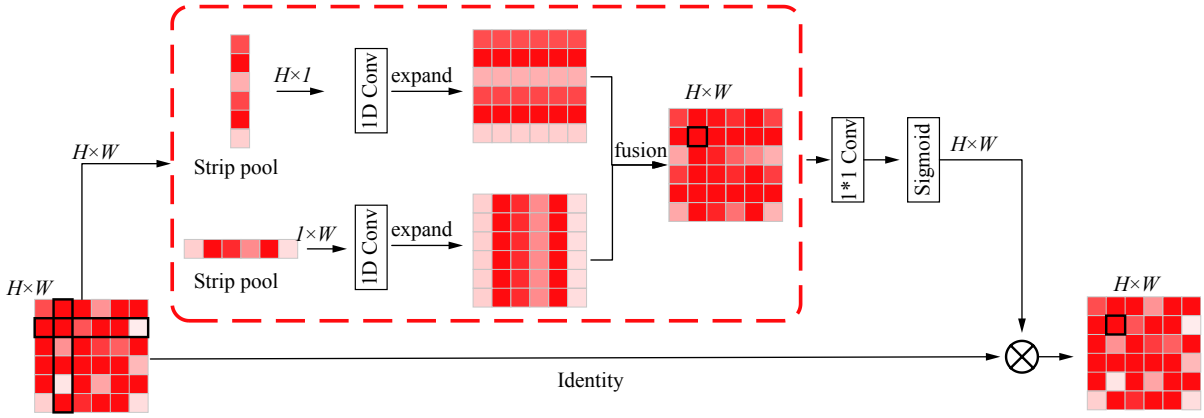


图 2　条形池化模块的网络框架

Fig. 2　Network framework of strip pooling module

该模块的具体工作过程可分为三步：(1)对输入的一个 $H\times W$ 的特征图，先在横向和纵向区域分别进行平均池化，池化核大小分别为 $W\times 1$ 和 $1\times H$ 并分别得到 2 个向量，向量尺寸为 $H\times 1$ 和 $1\times W$，然后再通过一维的卷积建立相邻区域的联系。(2)采用复制的方式对特征图进行横向和纵向扩张，然后再进行融合，融合方式为直接逐元素相加。(3)采用 1×1 的卷积对整个特征图进行变换，然后再通过 Sigmoid 函数进行权重归一化，最后分配到各个空间位置并与其进行相乘。

1.3　高效通道注意力模块

高效通道注意力模块[8]是在著名的通道注意力模块(即：SE 模块)[3]的基础之上改进得到的模块。研究发现，普通的 SE 模块存在 2 个方面的问题：首先，SE 模块虽然是轻量化的模块，可是参数量还是比较大；其次，传统的 SE 模块在通道转换部分是 2 个全连接层，这样做可以节省计算量和参数量，可同时也破坏了原有的权重和通道之间的对应关系。高效通道注意力模块是传统 SE 模块的改进版，如图 3 所示。

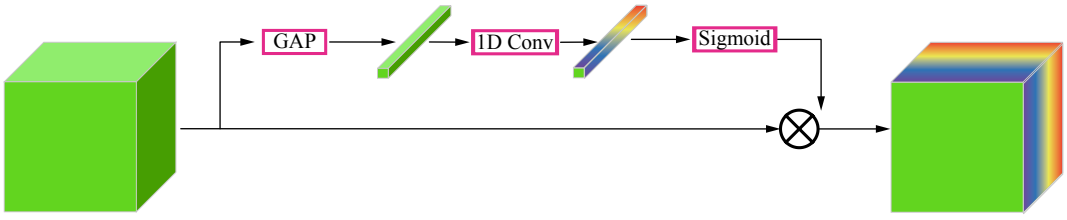


图 3　高效通道注意力模块的网络框架

Fig. 3　Network framework of efficient channel attention module

首先，对输入的特征图进行全局平均池化，其目的是对每个通道的压缩信息进行压缩，得到一个向量。其次，通过一个一维的 same 卷积进行处理，得到一个相同维度的向量，而这个卷积操作，其实就已经建立了每个通道及其相邻通道之间的函数关系。再次，通过 Sigmoid 函数在对数据进行数值归一化的同时增加网络的非线性。最后，再作为权重分配到各个通道，与各个通道的特征相乘之后得到最后的输出。

2　实　验

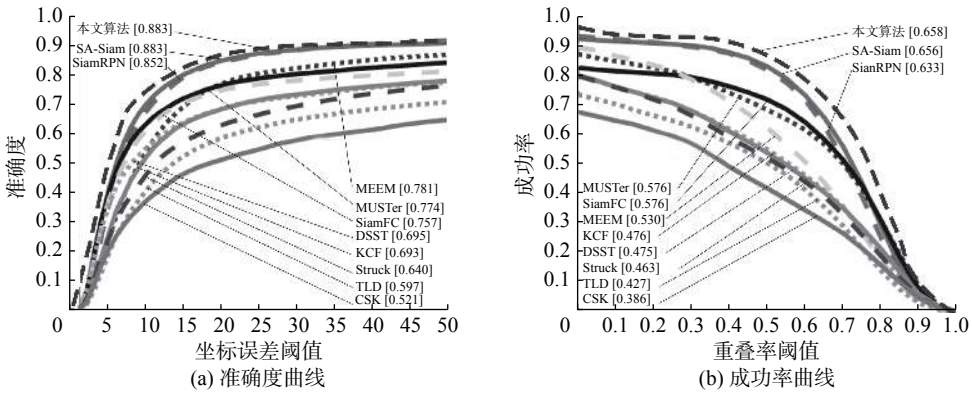
2.1　实验细节

本文把 ImageNet[13] 预先训练的 AlexNet[12] 作为骨干网络，共训练 20 个 epoch。先将骨干网络的参数固定，训练其他部分，训练 10 个 epoch 后，解除骨干网络后两层的冻结，并将其和网络的其他部分一起训练。

训练时，将模板帧的图像大小调整为 255×255 个像素点，搜索帧的图像大小调整为 127×127 个像素点。为了得到更好的训练效果，将 COCO[14]、Youtube-BB[15]、ImageNet VID 和 ImageNet DET4 个数据集作为训练集。并采用随机梯度下降法(SGD)进行训练。前 5 个 epoch 仅仅训练区域提议网络(RPN)部分。学习率从 0.005 均匀增加到 0.010。在随后的 25 个 epoch 中，整个网络的端到端训练的学习速率呈指数衰减，从 0.010 0 衰减到 0.000 5。使用 0.000 5 的重量衰减和 0.9 的动量。训练总损失是分类损失与回归的标准平滑 L_1 损失之和。本文实验使用 PyTorch 框架，硬件采用了 Intel(R)Xeon(R) CPU E5-1620 v3 @3.50 GHz，2 台英伟达 GTX 1080Ti GPU，内存 19 GB。

2.2　数据集与实验结果分析

使用标准的 OTB100[9] 基准和 100 个视频序列来评估本文提出的跟踪算法性能。对此前的 OTB2013 数据集[16] 进行了进一步的扩增，这些序列共有 11 种类型挑战，即：光照变化(IV)、变形(DEF)、运动模糊(MB)、平面外旋转(OPR)、低分辨率(LR)、遮挡(OCC)、快速运动(FM)、平面内旋转(IPR)、视野消失(OV)、背景混乱(BC)和尺度变化(SV)。评估指标有 2 个，分别为预测框与标准框的交并比(即成功率)和中心定位误差(即准确度)，图 4 为 10 个常见方法与本文方法的准确度和成功率曲线。



(a) 准确度曲线　　(b) 成功率曲线

图 4　不同算法在 OTB100 上的结果

Fig. 4　Results of different algorithms on OTB100

在图 4 中，精确度的纵坐标显示了满足中心定位要求的帧所占的百分比，成功率的纵坐标显示了满足该重叠率的帧所占的百分比。经本文与其他 10 个常见的方法相比，可以得到：本文的跟踪算法取得了很好的跟踪效果，无论是在成功率还是在准确度方面，都取得了第 1 名的好成绩；与基准算法 SiamRPN[17] 相比，本文算法的成功率提高了 3.1 个百分点，准确度提高了 2.5 个百分点，提升较为明显。

图 5 为本文算法与经典的 2 个目标跟踪算法 SiamRPN[17] 和 SiamFC[18] 在 OTB100[9] 数据集的 CliBar、Woman、DragonBaby、Coke、Jump 和 Matrix 6 个视频序列的部分视频帧的跟踪效果。从图 5 可以看出：本文提出的算法，无论是在定位，还是在尺度估计上，都明显优于其他 2 个跟踪算法；特别是当物体出现较快的位移(第 1 行的杂志和第 3 行的小孩)或者部分遮挡(第 2 行的行人和第 4 行的可乐瓶)时，本文的跟踪算法仍然可以保持良好的跟踪性能。

在 VOT2016[10] 数据集上做了测试，并且与

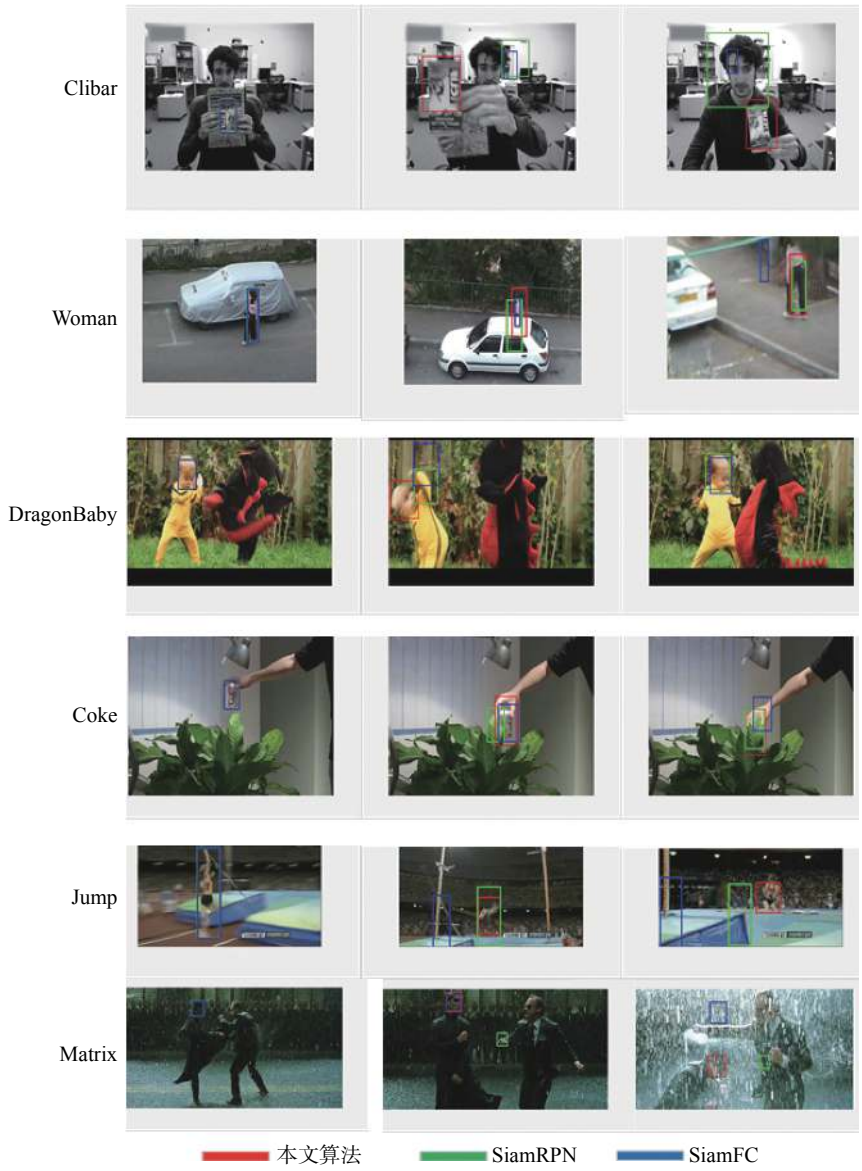


图 5　不同算法在 OTB100 上的跟踪效果

Fig. 5　Tracking effect of different tracking algorithms on OTB100

最先进的 9 个跟踪算法做了比较。VOT2016 公开数据集用于单目标的短期跟踪，其中包含 60 个视频序列。采用 Expected Average Overlap（*EAO*），Accuracy（*A*）和 Robustness（*R*）3 个指标进行比较不同的跟踪器，*A*、*R* 结果如表 1 所示，*EAO* 结果如图 6 所示。

从表 1 和图 6 可以看出：本文算法的 *EAO*、*A* 和 *R* 都处于第 2 名的位置（其中①，②和③分别代表第 1、第 2 和第 3 名，*EAO*、*A* 数值越高性能越好，*R* 数值越低性能越好），*EAO* 和 *A* 仅仅比第 1 名低了 0.2 和 0.1 个百分点，鲁棒性也只差 0.4 个百分点；总体效果上，排名第 1 的 DaSiamRPN[19] 跟踪算法，虽然在准确度上取得了领先，但是由于其采用了全局搜索操作影响了跟踪的速度，因而在速度方面低于本文算法。综合以上考虑，本文的跟踪算法取得了较为良好的效果。

VOT2019[11] 是在 VOT2018[20] 的基础上改进的，替换了其中的部分序列，视频总数依然是 60 个，仍然采取 *EAO*、*A* 和 *R* 3 个指标进行不同跟踪算法的比较。*A*、*B* 比较结果如表 2 所示，*EAO* 结果如图 7 所示。

表 1　不同跟踪算法在 VOT2016 上的结果
Tab. 1　Results of different tracking algorithms on VOT2016

算法	*A*	*R*
DaSiamRPN	①0.610	①0.220
Ours	②0.609	②0.224
SiamDW	③0.580	0.240
SiamRPN	0.560	0.260
CCOT	0.539	0.238
TCNN	0.554	0.268
SSAT	0.577	0.291
MLDF	0.490	③0.233
Staple	0.544	0.378
EBT	0.465	0.251

注：①、②、③分别表示第1名、第2名、第3名

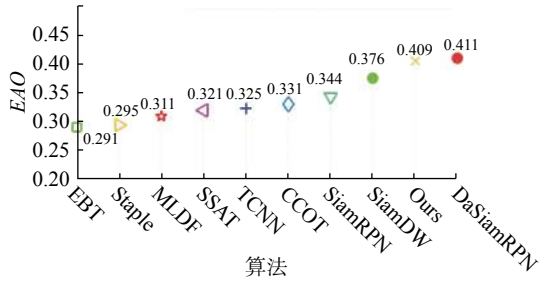


图 6　不同跟踪算法在 VOT2016 的 *EAO*

Fig. 6　*EAO* of different tracking algorithms in VOT2016

表 2　不同跟踪算法在 VOT2019 上的结果
Tab. 2　Results of different tracking algorithms on VOT2019

算法	*A*	*R*
Ours	①0.564	①0.495
SSRCCOT	0.495	②0.507
MemDTC	0.485	0.587
SiamRPNX	0.517	③0.552
Siamfcos	②0.561	0.788
TADT	③0.516	0.677
CSRDCF	0.496	0.632
CSRpp	0.468	0.662
FSC2F	0.480	0.752
ALTO	0.358	0.818

注：①、②、③分别表示第1名、第2名、第3名

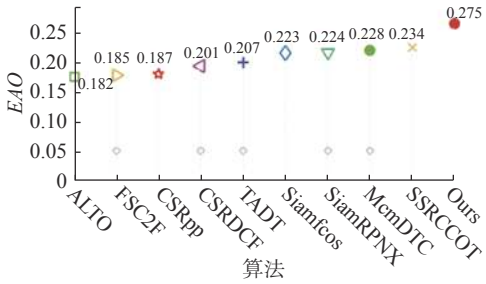


图 7　不同跟踪算法在 VOT2019 的 *EAO*

Fig. 7　*EAO* of different tracking algorithms in VOT2019

从表 2 和图 7 可以看出：本文的 *EAO*、*A* 和 *R* 都处于第一名的位置（其中①，②和③分别代表第 1，第 2 和第 3 名，*EAO*、*A* 数值越高性能越好，*R* 数值越低性能越好），*EAO* 和 *A* 分别比第 2 名的算法高了 4.1 和 6.9 个百分点，鲁棒性也好了 1.2 个百分点；总体效果上，比第 2 名的跟踪算法优秀很多，比经典的跟踪算法 SiamRPNX 也高出很多。

3　结束语

针对目标跟踪中物体长宽比比较悬殊和有干扰信息的问题，分别加入了条形池化模块和高效通道注意力模块，使得该问题得到了较好的解决，而且都是轻量级模块，对网络的推理速度影响可以忽略不计，且很容易训练，后续的实验也充分证明了，这样的改进，对于原本跟踪算法提升精度较为明显。

参考文献：

[1] HOU Q B, ZHANG L, CHENG M M, et al. Strip pooling: rethinking spatial pooling for scene parsing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE, 2020.

[2] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems. Cambridge: ACM, 2015: 91 − 99.

[3] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern

Analysis and Machine Intelligence, 2020, 42(8): 2011 – 2023.

[4] LI X, WANG W H, HU X L, et al. Selective kernel networks[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 510 − 519.

[5] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//15th European Conference on Computer Vision. Munich: Springer, 2018: 3 – 19.

[6] WANG X L, GIRSHICK R, GUPTA A, et al. Non-local neural networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 7794 – 7803.

[7] SZEGEDY C, LIU W, JIA Y Q, et al. Going deeper with convolutions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Boston: IEEE, 2015: 1 – 9.

[8] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[J]. arXiv preprint arXiv: 1910.03151, 2019.

[9] WU Y, LIM J, YANG M H. Object tracking benchmark[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(9): 1834 – 1848.

[10] KRISTAN M, LEONARDIS A, MATAS J, et al. The visual object tracking VOT2016 challenge results[C]// European Conference on Computer Vision. Amsterdam: Springer, 2016: 777 – 823.

[11] KRISTAN M, MATAS J, LEONARDIS A, et al. The seventh visual object tracking VOT2019 challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea: IEEE, 2019: 2206 – 2241.

[12] KRIZHEVSKY A, SUTSKEVER I, HINTON G E, et al. ImageNet classification with deep convolutional neural networks[C]//Proceedings of the 25th International Conference on Neural Information Processing Systems. Red Hook: ACM, 2012: 1097 – 1105.

[13] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211 – 252.

[14] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]//13th European Conference on Computer Vision. Zurich: Springer, 2014: 740 – 755.

[15] REAL E, SHLENS J, MAZZOCCHI S, et al. YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 7464 – 7473.

[16] WU Y, LIM J, YANG M H, et al. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 2411 – 2418.

[17] LI B, YAN J J, WU W, et al. High performance visual tracking with siamese region proposal network[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8971 – 8980.

[18] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]//European Conference on Computer Vision. Amsterdam: Springer, 2016: 850 − 865.

[19] ZHU Z, WANG Q, LI B, et al. Distractor-aware siamese networks for visual object tracking[C]//15th European Conference on Computer Vision. Munich: Springer, 2018: 103 – 119.

[20] KRISTAN M, LEONARDIS A, MATAS J, et al. The sixth visual object tracking VOT2018 challenge results[C]//European Conference on Computer Vision. Munich: Springer, 2018: 3 – 53.

（编辑：刘铁英）