中国科技期刊卓越行动计划项目入选期刊

控制与决策

CONTROL AND DECISION

融合多模板注意力深度网的自适应目标框跟踪算法

仲训杲, 范东嘉, 仲训昱, 周承仙, 赵晶, 刘强

引用本文:
仲训杲,范东嘉,仲训昱,周承仙,赵晶,刘强. 融合多模板注意力深度网的自适应目标框跟踪算法[J]. 控制与决策, 2024, 39(4): 1123−1132.

在线阅读 View online: https://doi.org/10.13195/j.kzyjc.2022.1670

融合多模板注意力深度网的自适应目标框跟踪算法

仲训杲[1,4†], 范东嘉[1], 仲训昱[2], 周承仙[1], 赵 晶[1,4], 刘 强[3]

(1. 厦门理工学院 电气工程与自动化学院，福建 厦门 361024；
2. 厦门大学 航空航天学院，福建 厦门 361002；
3. 牛津大学 精神学系，牛津 OX3 7JX；
4. 厦门市高端电力装备及智能控制重点实验室，福建 厦门 361024)

摘 要: 现有深度网络跟踪算法应对相似物体干扰、尺度变化、形变模糊、遮挡等问题存在挑战,为此提出一种融合多模板注意力机制的鲁棒深度网络算法. 在SiamFc深度网络分支中构建通道和空间多模板注意力机制,以加强网络对目标特征的提取能力;融合浅层和深层卷积特征实现跟踪目标的精确聚焦,以克服相似物干扰问题;采用自适应回归网络学习目标采样点与目标边界之间的距离,实现目标区域的动态预测,有效应对目标尺度变化问题. 另外,通过计算分类特征的APCE均值和最大值建立模板在线更新策略,实现网络自适应目标形变模糊与遮挡等问题. 对OTB 100和VOT 2016等公开数据集的测试结果表明,与目前先进的SiamFc及改进方法相比,所提出算法在动态目标跟踪的准确率和成功率上均得到有效提升,具有强鲁棒性能.

关键词: 深度网络；目标跟踪；自适应框；注意力机制；模板更新

中图分类号: TP391.4 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2022.1670

引用格式: 仲训杲,范东嘉,仲训昱,等. 融合多模板注意力深度网的自适应目标框跟踪算法[J]. 控制与决策, 2024, 39(4): 1123-1132.

Adaptive target box tracking algorithm by integrating multi-template attention deep network

ZHONG Xun-gao[1,4†], *FAN Dong-jia*[1], *ZHONG Xun-yu*[2], *ZHOU Cheng-xian*[1], *ZHAO Jing*[1,4], *LIU Qiang*[3]

(1. School of Electrical Enginnering and Automation，Xiamen University of Technology，Xiamen 361024，China；2. School of Aerospace Engineering，Xiamen University，Xiamen 361002，China；3. Department of Psychiatry，University of Oxford，Oxford OX3 7JX，UK；4. Xiamen Key Laboratory of Frontier Electric Power Equipment and Intelligent Control，Xiamen 361024，China)

Abstract: In view of the similar object interference, target scale changes, deformation blur, occlusion and other challenging problems for existing deep network tracking algorithms. This paper proposes a robust deep network tracking algorithm by integrating multi-template attention mechanism. The proposed method builds a channel and spatial multi-template attention mechanism in the branch of the Siamfc network, so as to strengthen the ability of the deep network for features extraction, and by integrates shallow and deep convolution features to achieve the accurate focus of tracking targets, so as to overcome the interference problem of similar objects. The adaptive regression network is used to learn the distance between the target sampling point and the target boundary, so as to realize the dynamic prediction of the target area and effectively deal with the problem of target scale change. In addition, the target template online update strategy is established by calculating the APCE mean value and maximum value of classification features, so as to realize the network adaptive the target deformation blur and occlusion problems. Through the test of OTB100, VOT2016 and other public data sets, the results show that compared with the current advanced deep network frameworks such as Siamfc and its improved method, the proposed algorithm has effectively improved the accuracy and success rate of dynamic target tracking, and the research method has a strong robust performance.

Keywords: deep network；object tracking；adaptive box；attention mechanism；template update

收稿日期: 2022-09-21；**录用日期:** 2022-12-08.

基金项目: 国家自然科学基金项目(61703356)；福建省自然科学基金项目(2022J011256, 2020J01285)；厦门市青年创新基金项目(3502Z20206071).

责任编委: 巩敦卫.

†通讯作者. E-mail: zhongxungao@163.com.

0 引 言

动态目标的鲁棒跟踪是计算机视觉长期研究的技术问题[1],针对诸如目标形变模糊,尺度变化,相似物体干扰,遮挡等挑战,提出全新的鲁棒跟踪算法成为当前视觉跟踪的研究新热点.

目前主流目标跟踪算法分为相关滤波算法和深度学习算法. 相关滤波算法的核心思想是基于模板匹配技术,代表性研究成果MOSSE (minimum output sum of squared error filter)算法[2]将相关滤波思想用于目标跟踪,该算法因跟踪速度表现不俗而得到广泛关注与改进. 如:针对MOSSE只使用单一的灰度特征问题,KCF (kernelized correlation filters)算法[3]使用多通道特征使目标表达更全面. 文献[4]对颜色特征空间进行扩展,然后将各特征进行加权,以有效提升算法的跟踪精度. Staple算法[5]同时使用HOG和CN特征,在一定程度上增强了视觉特征在目标形变上的鲁棒性.

深度神经网络能有效提升目标表达的鲁棒性. 文献[6]基于相似度匹配提出了全卷积孪生网络(fully-convolutional siamese networks, SiamFc)目标跟踪算法,开创了深度神经网络用于目标跟踪的一个重要研究分支. 文献[7]在孪生SiamFc网络中加入相关滤波器,进一步强化了网络对特征的学习能力. 为了使孪生网络在多尺度上更好地拟合目标,文献[8]采用区域候选网络(RPN)对预测框进行分类和回归. SiamRPN++网络[9]使用更深的网络结构进行特征提取,使目标表达更加稳定. SiamBAN算法[10]使用无锚框策略以适应多尺度目标. SiamRN算法[11]引入细化模块和关系检测器两个网络结构,在一定程度上提高了算法的跟踪准确率.

SiamFc算法使用AlexNet[12]作为主干网络对模板图片和搜索图片进行特征学习,对得到的多尺度特征无差别利用,而未考虑特征提取过程中不同空间信息和通道信息对目标表达的不同作用;SiamFc算法使用主干网络深层特征作为目标特征,忽略了浅层特征对目标的结构性表达;SiamFc使用尺度缩放因子预测目标框尺寸,由于尺度缩放因子数量的局限导致SiamFc在面对目标尺度变化时难以精确拟合目标框,同时,使用互相关运算在面对相似物时容易对背景产生较高响应;SiamFc算法只使用视频序列中的第1帧目标作为模板,对后续帧目标变化缺乏自适应性.

视觉注意力最早在图像处理领域开展应用研究,文献[13]提出压缩激励网络(SENet)在通道维度对重要视觉信息进行关注. 文献[14]提出了ECANet考虑k个相邻通道捕获局部跨通道的交互信息,建立不降维通道注意力模型. 文献[15]针对卷积感受野局限性,提出了(non-local neural networks NLNet)在空间维度提高重要信息的关注度. CCNet模型[16]通过计算每个像素与其同行同列像素的相似性,实现间接计算像素之间的相似性. CA (coordinate attention)模型[17]同时考虑通道关系和目标位置信息,取得了目标注意的显著成效.

综上,针对SiamFc及改进算法存在待解决的问题,并结合注意力模型的优势,本文提出融合多模板注意力机制的自适应目标跟踪新算法,旨在提升全卷积孪生网络对目标跟踪的鲁棒性能. 为此,本文主要研究内容包括以下4个方面:1) 考虑相似物体干扰、目标尺度变化、形变模糊、遮挡等挑战问题,设计通道和空间多模板注意力模型,加强有用通道和空间特征并抑制冗余特征,提升网络对目标的辨别能力;2) 将目标浅层空间特征和深层语义特征以残差方式进行有效融合,强化目标特征的有效表达;3) 使用深度互相关卷积改进传统互相关运算,并添加目标框自适应回归网对目标位置进行精确预测,提升算法的鲁棒性能;4) 使用APCE[18]阈值条件下的目标模板更新策略,加强模板对动态目标的自适应能力. 通过对公开数据集的对比测试,实验结果表明了所提出研究方法的鲁棒性能.

1 SiamFc孪生算法及问题分析

SiamFc算法的核心思想是通过比较模板图片和搜索图片的相似度,确定目标在搜索图片中的位置. 一般地,模板图预处理为127×127大小的尺度,而搜索图预处理为255×255尺度,将预处理图片输入权值共享的5层AlexNet孪生网络实现特征提取,之后对模板和搜索图特征做互相关操作,得到17×17×1尺度的响应图,计算方式[6]为:

$$e = f(Z, X) - f_{\varphi}(Z) * f_{\varphi}(X) + b. \tag{1}$$

其中:f_{φ}为卷积函数,Z和X为模板图片和搜索图片,$*$为互相关操作,b为偏置项,e为互相关响应图. 将响应图最大值位置映射到搜索图中实现目标跟踪. SiamFc使用logistic损失函数对样本进行训练,即

$$l(y, v) = \log(1 + \exp(-yv)). \tag{2}$$

其中:$y \subseteq (+1, -1)$为真值,v为搜索图像实际得分. 训练时采用所有候选位置的平均损失,并使用随机梯度下降法对网络参数进行优化,损失函数如下:

$$L(y,v)=\frac{1}{D}\sum_{u\subseteq D}l(y[u],v[u]). \tag{3}$$

其中: D 为得分响应图, u 为得分响应图中所有的位置. SiamFc方法将所有特征平等对待, 无用特征对跟踪产生干扰; 主干网络特征对目标表达层次不够全面; 有限尺度缩放无法应对目标任意尺度变化; 跟踪模板面对目标动态变化易丢失.

2　鲁棒深度网络框架设计

针对SiamFc孪生算法存在的上述问题, 本文提出鲁棒深度网络结构, 如图1所示特征提取主干网采用AlexNet孪生架构. 为了加强目标相关特征的作用力, 抑制目标无关特征的干扰, 设计通道和空间多模板注意力机制对目标特征进行加权, 使网络聚焦目标有用特征, 强化网络对目标的辨别能力. 为了弥补SiamFc只使用深层抽象特征对目标表达的不足, 将主干网络的浅层和深层卷积输出特征进行加权融合, 使网络对目标的表征更为全面. 针对目标运动过程尺度变化问题, 使用目标框自适应回归网络对深度互相关特征进一步学习, 得到判断目标位置的分类特征图和拟合目标的回归特征图, 再通过预测目标采样点与目标边界的距离得到目标预测框. 考虑到目标动态跟踪过程后续目标形态与目标模板差异较大, 将通过计算分类分支的APCE值更新目标模板, 加强模板对动态目标的自适应能力.

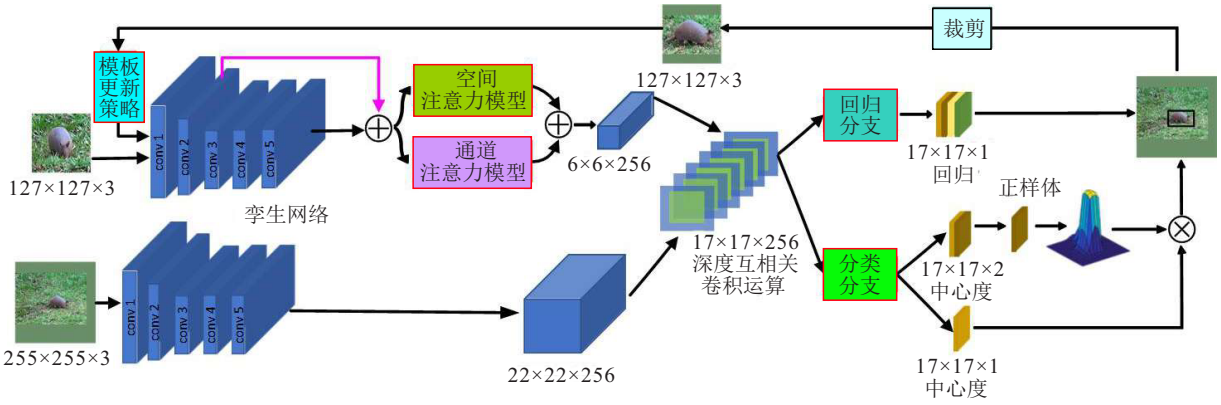


图 1　鲁棒深度网络结构

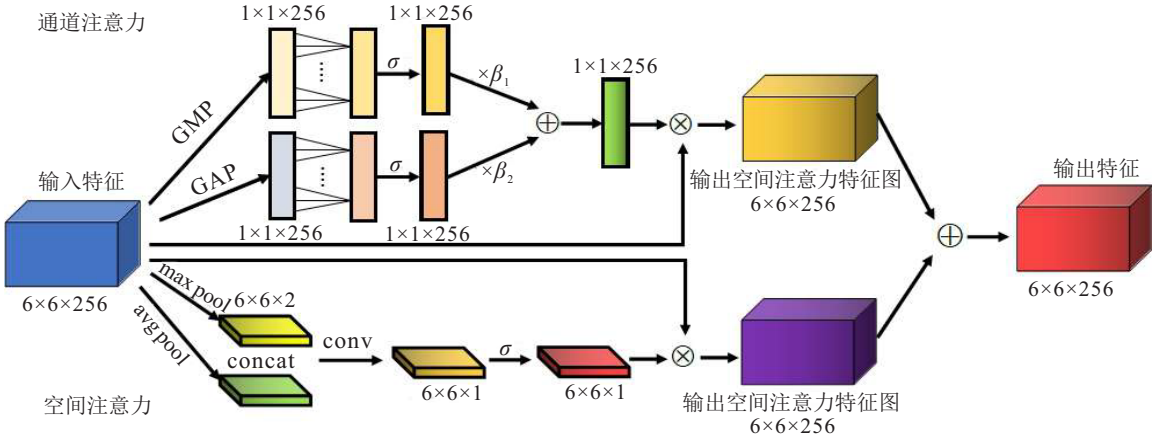


图 2　多模板注意力模型

2.1　多模板注意力模型设计

如图1所示, 注意力模型作用于目标特征提取网络, 在通道和空间维度上加强有用特征信息, 抑制冗余信息的不良影响. 本文所设计注意力模型与CBAM模型[19]的区别体现在模型结构和融合方式不同. 本文设计的通道注意力网络在完成特征图通道做全局最大池化和全局平均池化操作得到两个特征图之后, 经过跨通道交互对特征进行学习, 相比于CBAM模型, 本文方法避免了降维, 减少了对通道注意力的预测产生的负面影响. 本文多模板注意力模型如图2所示, 包括通道注意力模型和空间注意力模型, 其中通道注意力模型以主干网络的输出特征作为输入. 为了获得特征图中每个通道最显著性的特征, 使用全局最大池化(GMP)作用于特征图通道, 同时, 使用全局平均池化(GAP)获得各通道的平均特征, 将两个输出结果分别进行通道间的卷积学习, 获得相邻通道间的相互依赖性, 之后使用Sigmoid函数进行非线性学习得到各分支的通道权重, 权重计算公式为

$$b'_c = \sigma\Big(\sum_{j=1}^{k} a^j b_c^j\Big),\ b_c^j \subseteq \Omega_c^k. \tag{4}$$

其中: Ω_c^k 表示 b_c 的 k 个相邻通道的集合, k 设置为3, σ 为Sigmoid激活函数, b'_c 为通道权重,包括最大池化和平均池化分支权重,通过加权融合权衡两种权重的作用为

$$B_c = \beta_1 b'_{c\,\max} \oplus \beta_2 b'_{c\,\mathrm{avg}}. \tag{5}$$

其中: $b'_{c\,\max}$ 和 $b'_{c\,\mathrm{avg}}$ 分别为最大池化和平均池化得到的通道权重; β_1 和 β_2 为融合权重系数,考虑不同场景的跟踪性能,设置融合权重系数实现均衡两种注意力的作用,因此本文取0.5; B_c 为融合后的通道权重,最后将各通道权重赋予输入特征图对应通道.

空间注意力模型的作用是为不同空间位置特征分配不同的权重,因此本文构建的空间注意力机制与CBAM[19]级联方式不相同,而是将通道和空间两种注意力机制采用权重融合的方式. 首先,对空间位置特征执行全局最大池化,获得对目标表达最显著的特征;同时,使用全局平均池化获得空间位置中的平均特征;之后,将两种特征在通道方向上拼接,并通过卷积对拼接特征进一步学习;最后,使用Sigmoid函数作用获得空间注意力权重,即

$$b_s = \sigma(f^{3\times3}(\mathrm{concat}([\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)]))). \tag{6}$$

其中: σ 为sigmoid激活函数, $f^{3\times3}$ 为 3×3 卷积, b_s 为学习到的空间权重.最后将空间权重赋予输入特征中对应的空间位置得到空间注意力特征图. 将得到的通道和空间注意力特征图加权融合,同样为了均衡两种注意力机制的作用. 本文设置通道和空间注意力特征图的权重系数 λ_1、λ_2 均为0.5.

2.2 特征残差融合算法

图1中,随着神经网络前向传播,得到的特征分辨率越来越低,所包含的空间结构细节信息不断减少,目标特征变得越来越稀疏抽象. 因此,本文将深浅两种特征进行有效融合,使目标的特征表达更加全面. 具体地,对第2层卷积特征使用最近邻插值下采样后,与第5层卷积特征以残差连接的方式进行线性融合,其中最近邻插值法为

$$\begin{cases} \mathrm{src}\,X = \mathrm{dst}\,X * \left(\dfrac{\mathrm{src}\,W}{\mathrm{dst}\,W}\right); \\ \mathrm{src}\,Y = \mathrm{dst}\,Y * \left(\dfrac{\mathrm{src}\,H}{\mathrm{dst}\,H}\right). \end{cases} \tag{7}$$

其中: $\mathrm{dst}\,X$ 和 $\mathrm{dst}\,Y$ 为目标图像像素的横、纵坐标; $\mathrm{dst}\,W$ 和 $\mathrm{dst}\,H$ 为目标图像的宽、高; $\mathrm{src}\,W$ 和 $\mathrm{src}\,H$ 为原图像的宽、高; $\mathrm{src}\,X$ 和 $\mathrm{src}\,Y$ 为目标图像在该点 $(\mathrm{dst}\,X, \mathrm{dst}\,Y)$ 对应的原图像坐标. 图3为特征融合结果,既保留了目标浅层特征部分结构化信息,又具有深层特征抽象语义属性.

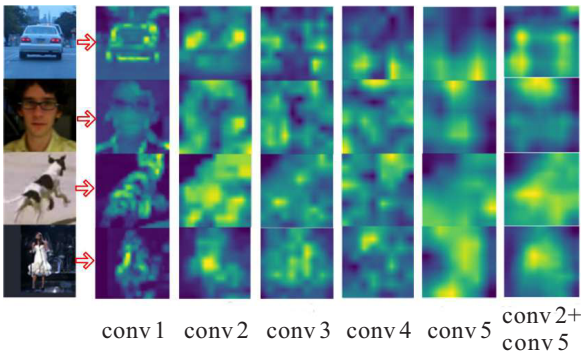


图 3 不同卷积层特征和融合特征

2.3 目标框自适应回归网络

针对SiamFc使用有限尺度缩放应对目标尺度变化的不足,本文采用与siamRPN++[9]相似的深度互相关卷积运算对模板特征和搜索特征进行互相关操作,在此基础上添加目标框自适应回归网络对目标位置进行精确预测,提升算法的鲁棒性能. 在图1中,目标框自适应回归网络包含回归分支和分类分支. 其中,回归分支输出的特征图每个位置 (i,j) 都对应原图中一个感受野中心点 (x,y). 训练时将落入目标区域的点作为正样本,并预测该点到目标四周边界的距离 $g(x,y) = (l,t,r,b)$,同时将该点到目标边界的真实距离定义为 $g^*(x,y) = (l^*,t^*,r^*,b^*)$,使用IOU损失作为回归分支损失函数,即

$$L_{\mathrm{reg}} = \frac{1}{\sum I(g^*_{(i,j)})} \sum_{(i,j)} I(g^*_{(i,j)}) L_{\mathrm{IOU}}(g_{(x,y)}, g^*_{(x,y)}). \tag{8}$$

其中

$$I(g^*_{(i,j)}) = \begin{cases} 1,\ g^{*(k)}_{(i,j)} > 0, k = 0,1,2,3; \\ 0,\ \text{otherwise}. \end{cases} \tag{9}$$

如果 (i,j) 对应原图中的点 (x,y) 在目标区域内,则为1,否则为0. 分类分支输出的特征图用来区分目标和背景,实验发现所有正样本点距离真实目标中心位置越远其预测回归的目标框越差. 为此,使用中心度分支来惩罚距离目标中心位置远的点,如图1所示,中心度分支输出一个通道特征图,其中任意位置的中心度计算公式为

$$C(i,j) = I(g^*_{(i,j)}) \times \sqrt{\frac{\min(l^*,r^*)}{\max(l^*,r^*)} \times \frac{\min(t^*,b^*)}{\max(t^*,b^*)}}. \tag{10}$$

由式(10)可知,距离目标中心位置越远的像素点其中心度得分越低,越靠近目标中心位置其中心度得

分越高,可将中心度得分特征图与分类特征图对应位置相乘,会对远距离点进行抑制. 中心度分支的损失函数为

$$L_{\text{cen}} = \frac{-1}{\sum I(g^*_{(i,j)})} \sum_{I(g^*_{(i,j)})=1} C(i,j) * \log A(i,j) + (1 - C(i,j)) * \log(1 - A(i,j)), \tag{11}$$

其中 $A(i,j)$ 为中心度特征图 (i,j) 位置处的得分值. 分类损失 L_{cls} 使用交叉熵损失,所以总的损失函数为

$$L = L_{\text{cls}} + \alpha_1 L_{\text{cen}} + \alpha_2 L_{\text{reg}}, \tag{12}$$

其中 α_1, α_2 为权重系数,由实验测试设置为1和3.

2.4　模板更新机制

为了使模板特征能够自适应目标形变模糊、旋转、遮挡等问题,本文计算分类特征图的APCE值[18]最大得分值,采用APCE值对模板特征在线更新,其中APCE计算方式为

$$\text{APCE} = \frac{|F_{\max} - F_{\min}|^2}{\text{mean}\left(\sum_{w,h}(F_{w,h} - F_{\min})^2\right)}. \tag{13}$$

其中: $F_{\max}$ 和 $F_{\min}$ 分别表示分类特征图最大值和最小值, $F_{w,h}$ 为分类特征图任意位置的值. 为此,本文模板更新数学表达为

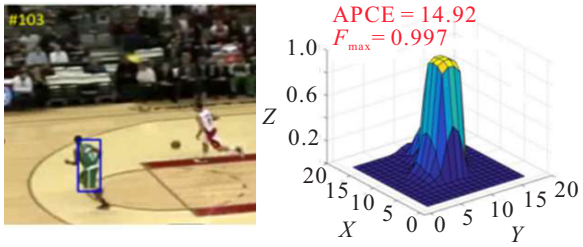
$$f_\varphi(z) = (1 - \eta) * f_\varphi(z^t) + \eta * f_\varphi(z^{t+u}). \tag{14}$$

其中: z^t 为基准模板图片, z^{t+u} 为更新模板图片, η 为更新系数,多次实验得到为0.01. 更新条件为

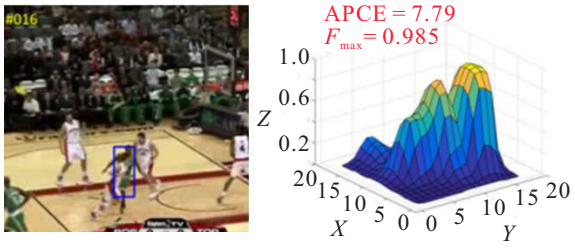
$$\begin{cases} \text{updata},\ P_{\text{cls}} > \upsilon \bar{P}_{\text{cls}},\ M_{\text{cls}} > \varpi \bar{M}_{\text{cls}}; \\ \text{no updata},\ \text{otherwise}. \end{cases} \tag{15}$$

其中: P 为APCE值, $\bar{P}$ 为APCE历史均值, M 为分类特征的最大得分值, $\bar{M}$ 为最大得分值历史均值, cls为分类特征图. 如果某一帧分类特征图APCE值大于APCE历史均值的 μ 倍,并且最大得分值大于最大得分值历史均值的 ϖ 倍时,算法将执行模板更新. 通过对不同场景的有限次跟踪测试得到 μ 为1.3, ϖ 为1时跟踪效果较理想.

与现有方法相比,本文模板更新的计算依据选择在分类分支正样本的位置,当分类分支的APCE值和最大值都以一定倍数大于APCE均值和最大值均值时对模板进行更新. 如图4(a)所示,随着目标的动态变化,当目标与背景差异明显且无遮挡时,目标位置得分应远高于其他位置,APCE值较大,此时更新模板特征;如图4(b)所示,在跟踪过程中面对目标遮挡或与背景相似,APCE值较小,此时不更新模板特征.



(a)　模板更新



(b)　模板不更新

图 4　模板更新策略

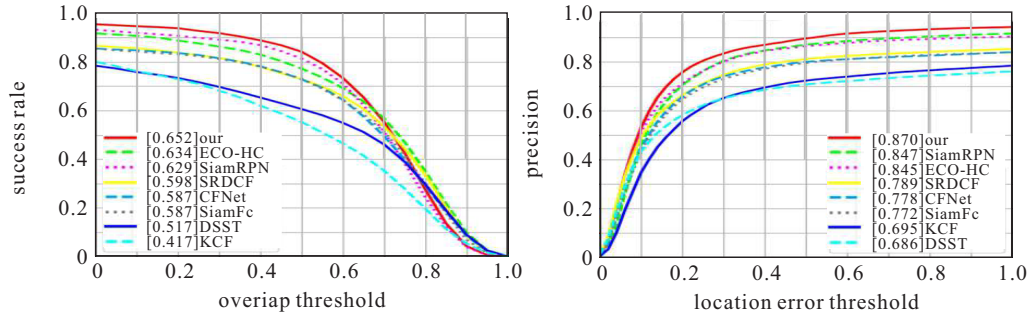
3　实验结果及分析

网络训练数据集采用Got-10 k[20],该数据集包含10 000个视频序列,563个目标类别,分为动物、人物、人造物体、自然物体以及其他5个大类. 网络模型共训练50个epoch,为了提升网络训练的高效性,前10个epoch冻结主干网络后两层参数,而后40个epoch对网络所有参数进行训练. 每批次训练图片数量为64,学习率初始化为0.01,前5个epoch将学习率增加至0.05,后45个epoch学习率使用指数衰减,使用随机梯度下降(SGD)对参数进行优化.

3.1　性能对比实验

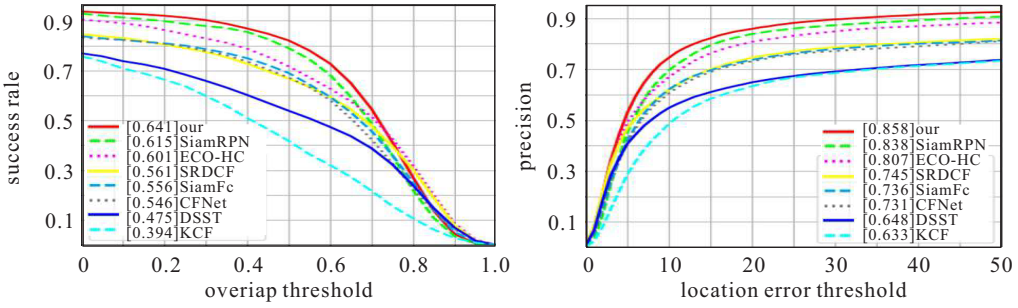
OTB 100实验: OTB 100[21]数据集包含100个跟踪视频序列,涵盖了跟踪任务中最常见的背景杂乱、目标形变、旋转、运动模糊等挑战性场景. OTB 100数据集采用跟踪成功率和精确率作为算法的评价指标,其中成功率定义为真实目标框与预测框重叠率大于阈值的帧数与总帧数的比值,而精确率定义为真实目标框中心与预测框中心距离小于阈值的帧数与总帧数的比值.

图5为本文算法在OTB 100数据集上与ECO-HC[22]、SRDCF[23]、DSST[24]等主流算法的比较结果. 可以看出,本文算法能够自适应拟合不同尺度的目标,并使用注意力机制加强目标的辨别能力,同时在线更新模板特征,最终成功率达到0.652,相较于基准算法SiamFc提升了6.5 %. 本文算法精确率为0.87,相较于基准算法SiamFc提升了9.8 %. 成功率和精确率相比于当前第2名的ECO-HC和SiamRPN 算法分别提升了1.8 %和2.3 %. 此外,与传统KCF等相关滤波算法相比具有显著优势.



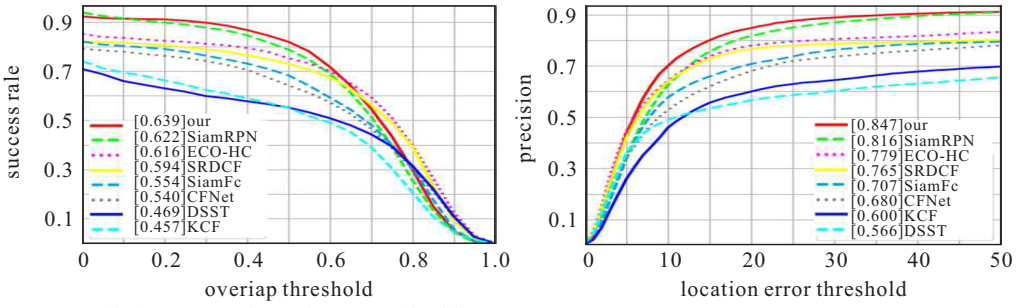
(a) success plots of OPE on OTB100 (b) precision plots of OPE on OTB100

图 5 不同算法在OTB 100数据集上的性能比较



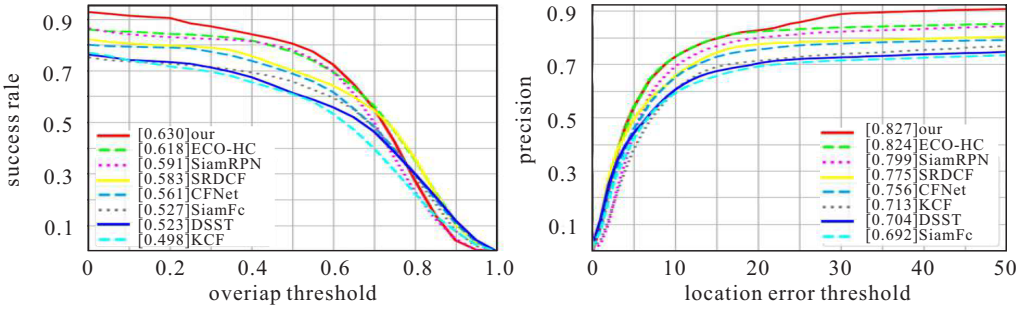
(a1) success plots of OPE-scale variation (a2) precision plots of OPE-scale variation

(a) 目标尺度变化场景测试结果



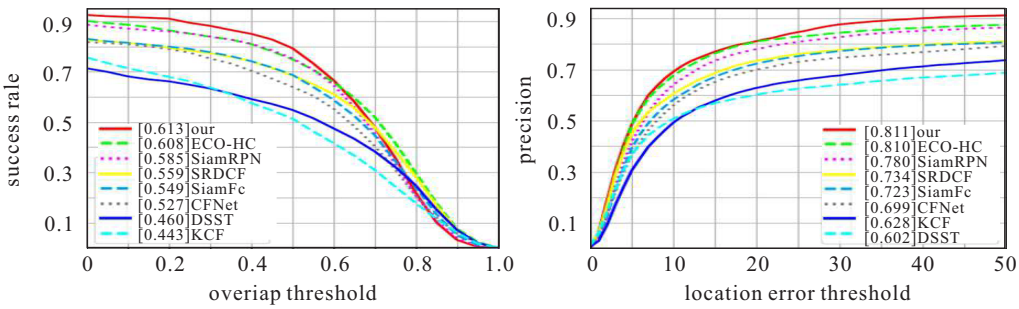
(b1) success plots of OPE-motion blur (b2) precision plots of OPE on OTB100

(b) 目标运动模糊场景测试结果



(c1) success plots of OPE-background clutters (c2) precision plots of OPE-background clutters

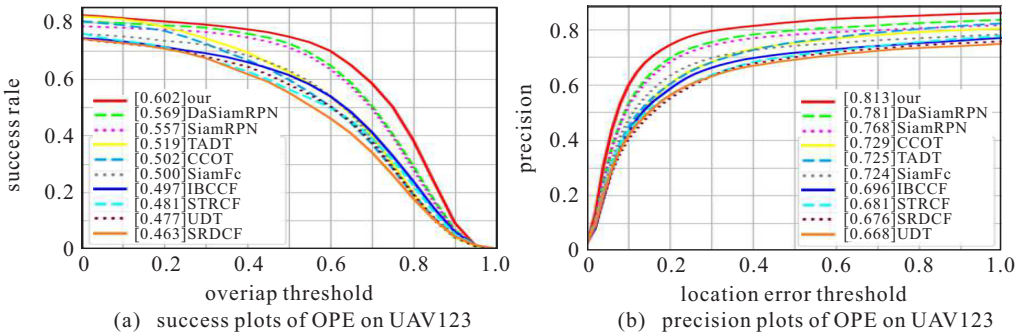
(c) 背景杂乱场景测试结果



(d1) success plots of OPE-occlusion (d2) precision plots of OPE-occlusion

(d) 目标遮挡场景测试结果

图 6 不同算法在OTB 100数据集不同场景下的性能比较



(a) success plots of OPE on UAV123　　(b) precision plots of OPE on UAV123

图 7　不同算法在UAV 123数据集上的性能比较

图 6 为 OTB 100 数据集不同场景下的测试结果, 其中图 6(a) 为目标尺度变化场景结果, 本文算法成功率和精确率为 0.641 和 0.858; 图 6(b) 为运动模糊场景结果, 本文算法成功率和精确率为 0.639 和 0.847; 图 6(c) 在背景杂乱场景下本文算法成功率和精确率为 0.630 和 0.827; 而图 6(d) 在遮挡场景下本文算法成功率和精确率为 0.613 和 0.811. 可以看出, 本文算法在不同条件下均取得了最优效果, 这主要得益于: 1) 多模板注意力模型和特征融合机制增强了算法对目标的辨别能力; 2) 目标框自适应回归网络能够拟合任意尺度目标; 3) 模板更新使模板特征能够自适应目标变化, 有效增强算法的鲁棒性.

UAV 123 实验: UAV 123 数据集[25] 由无人机拍摄得到, 包含 123 个视频序列, 视频帧中目标采取竖直矩形框进行标注. 与其他数据集相比, UAV 123 具有视角变化多的特点, 跟踪过程中目标会经历相机运动、尺度变化、背景杂乱和遮挡等复杂场景. 图 7 为本文算法和 DaSiamRPN[26]、TADT[27]、CCOT[28]、IBCCF[29]、STRCF[30]、UDT[31] 等多个代表性算法的性能评估结果. 可以看到, 本文算法成功率为 0.602, 精确率为 0.813, 相比于基准 SiamFc 算法成功率提高了 10.2 %, 精确率提高了 8.9 %, 而相比于排名第 2 的 DaSiamRPN 算法成功率提高了 3.3 %, 精确率提高了 3.2 %, 在所有对比算法中取得了最优的结果, 表明了本文提出的鲁棒深度网络架构应对复杂场景跟踪问题的有效性.

VOT 2016 实验: VOT 2016 数据集[32] 包含 60 个视频序列, 该数据集采用准确率 (accuracy)、鲁棒性 (robustness) 和期望平均覆盖率 (expected average overlap, EAO) 作为跟踪算法的评价指标. 其中准确率定义为预测框与真实框的交并比; 鲁棒性定义为跟丢帧数与总帧数的比值, 用来衡量跟踪算法的稳定性; 期望平均覆盖率评估算法的综合性能.

表 1 为本文算法与主流算法的对比结果, 本文算法准确率为 0.577, 鲁棒性为 0.196, 期望平均覆盖率为 0.383; 与 SaimFc 算法相比准确率提升 4.5 %, 鲁棒性提升 26.5 %, 期望平均覆盖率提升 14.8 %. 相比于排名第 2 的 SiamRPN 算法, 准确率提升 1.7 %, 鲁棒性提升 6.4 %, 期望平均覆盖率提升 3.9 %. 与传统 KCF、SRDCF 等相关滤波算法相比本文算法优势显著.

表 1　VOT 2016 数据集对比实验结果

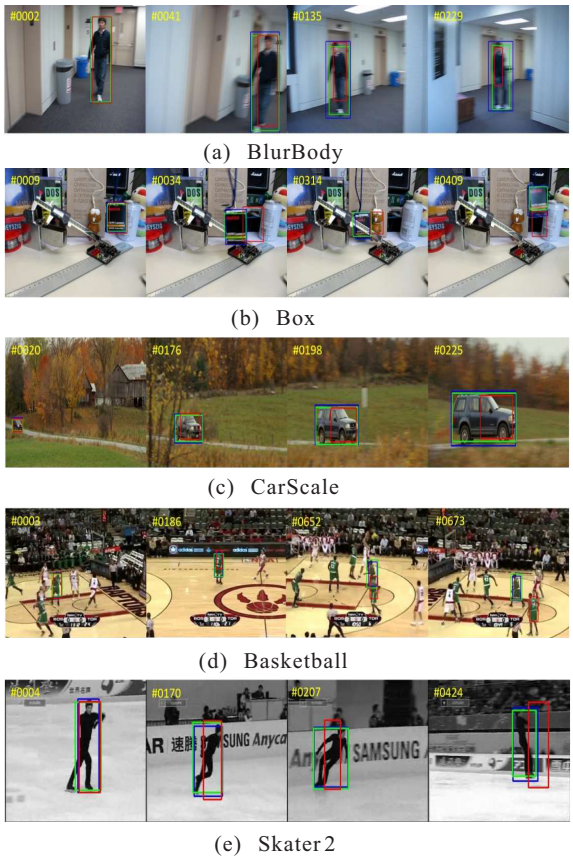
tracker	accuracy	robustness	EAO
DAT[33]	0.468	0.480	0.217
KCF[3]	0.489	0.569	0.192
ASMS[34]	0.503	0.522	0.212
SiamFc[6]	0.532	0.461	0.235
SRDCF[23]	0.535	0.419	0.247
Staple[5]	0.544	0.378	0.295
DensSiam[35]	0.560	0.330	0.331
SiamRPN[8]	0.560	0.260	0.344
Our	**0.577**	**0.196**	**0.383**

在 VOT 2016 数据集上测试本文算法运算速度可达到 73 fps, 与基准 SiamFc 算法 86 fps 具有一定的可比性, 主要因为本文注意力模型、目标框自适应回归网, 模板更新策略需要额外消耗一定的计算资源, 但相比于 SRDCF 算法 5 fps, 以及 DAT 算法 15 fps, 本文算法仍具有较大优势, 满足实时跟踪需求.

3.2　样本可视化实验

为清晰说明本文算法相比于基准 SiamFc 算法的优化效果, 本文可视化了 OTB 100 数据集典型挑战场景的跟踪结果, 测试结果如图 8 所示 (红色框为 SiamFc 算法, 绿色框为本文算法, 蓝色框为 Ground Truth), 具体分析如下.

1) 图 8(a) 的 BlurBody 序列中, 在第 1 帧和第 2 帧本文算法与基准 SiamFc 算法均能追踪目标. 在后续第 41 帧和第 135 帧目标发生运动模糊、形变和尺度变化, 本文算法生成的预测目标框较好地拟合了目标真实框, 而 SiamFc 预测框与真实框的重叠率不如本文算法. 在第 229 帧目标发生旋转, SiamFc 算法跟踪效果变得更差, 而本文算法依然能够较好地预测到目标位置.



(a) BlurBody

(b) Box

(c) CarScale

(d) Basketball

(e) Skater2

图 8 本文算法与基准 SiamFc 算法在 OTB 100 数据集上不同场景的跟踪结果

2) 在图 8(b) 的 Box 序列中, 在第 1 帧 ~ 第 9 帧 SiamFc 算法与本文算法均能很好追踪目标. 在第 34 帧和第 409 帧, 目标周围出现相似物, 此时 SiamFc 算法出现目标丢失现象, 而本文算法在相似背景场景中提高了跟踪成功率. 在第 314 帧, 目标被部分遮挡, 本文算法依然能够跟踪目标, 而 SiamFc 算法跟踪失败.

3) 在图 8(c) 的 CarScale 序列中, 在第 20 帧本文算法与 SiamFc 算法均能成功跟踪目标. 在第 176 帧目标发生旋转和尺度变化, 并且部分被遮挡, 可见 SiamFc 预测目标框不如本文算法. 在第 198 帧目标尺度变大, 并且发生旋转, 此时 SiamFc 只能识别车头部分, 而本文算法通过模板更新策略, 对目标尺度变化具有很好的自适应性. 在第 225 帧本文算法依旧能预测目标区域, 而 SiamFc 难以应对目标旋转、尺度变化问题.

4) 在图 8(d) 的 Basketball 序列中, 在起始帧本文算法与 SiamFc 算法均能对目标进行跟踪. 随着复杂场景的干扰, 如在第 186 帧, 本文算法目标预测框更接近真实框. 在第 652 帧和第 673 帧目标被部分遮挡, 并且背景存在相似物干扰, 本文算法依旧能正确识别并跟踪目标, 而基准 SiamFc 算法表现乏力.

5) 在图 8(e) 的 Skater2 序列中, 在第 170 帧和第 207 帧目标运动发生形变, 同时伴随旋转, 相比于基准 SiamFc 算法, 本文算法对目标的预测更佳. 在第 424 帧, 随着跟踪过程复杂场景的干扰, SiamFc 算法发生目标漂移, 而本文算法依然能够鲁棒追踪目标.

最后, 采用热度图直观展现注意力机制对目标的辨别能力和对背景信息的抑制作用. 结果如图 9 所示, 在 Basketball 和 DragonBaby 序列中, 面对相似物体干扰和运动模糊, 此时注意力机制能对相似背景信息进行有效抑制, 热度更聚焦于兴趣目标, 而非注意力产生了较为广泛的热响应, 不利于区分目标. 在 Bolt 序列中面对目标形变、复杂背景, 非注意力对背景区域产生大量响应, 形成多个热点, 而采用注意力机制模型响应更聚焦目标区域, 说明本文多模板注意力模型真实有效.

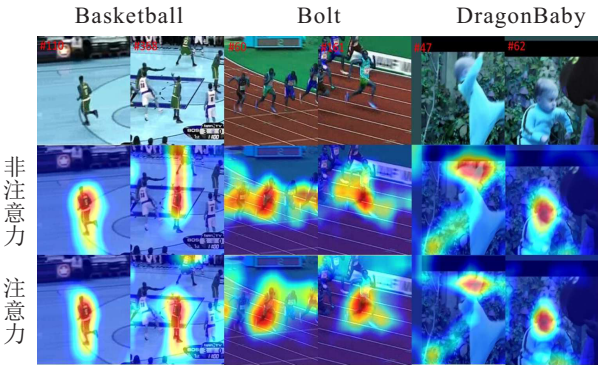


图 9 注意力机制目标聚焦结果

3.3 算法消融实验

为了更加直观地说明本文改进算法添加的残差融合模块、多模板通道注意力和空间注意力机制模块的有效性, 通过消融实验对各模块的作用进行测试. 表 2 为各模块在 OTB 100 数据集上精确率的消融实验结果. 其中: F 表示在 SiamFc 网络中增加残差融合模块, C 表示添加通道注意力机制模块, C+S 表示同时使用通道注意力机制和空间注意力机制模块, B 表示添加目标框自适应回归网, Our 是指在以上基础上再增加模板更新机制.

表 2 消融实验结果

对比模块	形变	模糊	杂乱	旋转	遮挡
SiamFc	0.691	0.707	0.692	0.743	0.723
F	0.744	0.749	0.723	0.782	0.733
C	0.751	0.733	0.732	0.773	0.735
C+S	0.762	0.751	0.739	0.7864	0.736
F+C+S	0.769	0.760	0.760	0.802	0.742
F+C+S+B	0.853	0.842	0.811	0.867	0.795
Our	**0.874**	**0.847**	**0.827**	**0.881**	**0.811**

从实结果可看出, 相比于 SiamFc 网络, 使用残差融合模块对目标形变、运动模糊、背景杂乱和旋转场景下目标跟踪的性能有较好的改善, 这是因为融合浅层特征的网络可以结合目标的轮廓、纹理、颜色等

特征对目标更好地识别. 使用通道注意力模块对表2中5种复杂场景下的目标跟踪精确率都有一定的提高,这是因为通道注意力机制强化有用通道信息的同时抑制无用通道信息,使网络对目标特征表达更清晰. 在通道注意力模块的基础上添加空间注意力模块,目标跟踪性能进一步提高,因为从空间维度凸显目标特征,有利于提升对目标的辨别.

在以上工作基础上使用目标框自适应回归网络,对目标和背景进行分类,同时改变目标框生成策略,对目标的辨别和目标框的拟合能力将进一步提高. 根据分类分支得分对目标模板进行更新,能够更好应对目标形变、背景杂乱、旋转和遮挡场景下的跟踪问题. 因此,通过消融实验可验证本文所提出的各个网络模块的有效性.

4 结 论

针对SiamFc孪生算法仅使用深层特征表达目标存在的片面性,将浅层特征与深层特征融合作为目标特征,增加结构性信息完善目标特征的表达;为了在特征提取过程中加强有用特征的作用而抑制无用特征的影响,设计通道和空间多模板注意力模型,对目标特征通道和空间位置进行加权;提出目标框自适应回归网络,改进SiamFc算法有限数量的尺度估计策略;同时根据分类特征计算APCE值实现对模板特征的在线更新. 结合以上改进策略,本文建立了基于SiamFc孪生算法的鲁棒深度网络框架,本文算法在OTB 100和VOT 2016等公开数据集上进行对比测试,结果表明本文算法在各项评价指标上均优于基准SiamFc算法和目前主流的目标跟踪算法,并验证了本文算法面对相似物体干扰、目标尺度变化、形变模糊、遮挡等挑战的鲁棒性能.

参考文献(References)

[1] Meng L, Yang X. A survey of object tracking algorithms[J]. Acta Automatica Sinica, 2019, 45(7): 1244-1260.

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

[3] 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.

[4] Henriques J F, Caseiro R, Martins P, et al. Exploiting the circulant structure of tracking-by-detection with kernels[C]. Proceedings of the 12th European conference on Computer Vision. New York, 2012: 702-715.

[5] 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 (CVPR). Las Vegas, 2016: 1401-1409.

[6] 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. Cham: Springer, 2016: 850-865.

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

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

[9] Li B, Wu W, Wang Q, et al. SiamRPN: Evolution of siamese visual tracking with very deep networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 4277-4286.

[10] Chen Z D, Zhong B N, Li G R, et al. Siamese box adaptive network for visual tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 6667-6676.

[11] Cheng S Y, Zhong B N, Li G R, et al. Learning to filter: Siamese relation network for robust tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 4419-4429.

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

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7132-7141.

[14] Wang Q L, Wu B G, Zhu P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, 2020: 11531-11539.

[15] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 7794-7803.

[16] Huang Z L, Wang X G, Huang L C, et al. CCNet: Criss-cross attention for semantic segmentation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, 2019: 603-612.

[17] Hou Q B, Zhou D Q, Feng J S. Coordinate attention for

efficient mobile network design[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, 2021: 13708-13717.

[18] Wang M M, Liu Y, Huang Z Y. Large margin object tracking with circulant feature maps[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 4800-4808.

[19] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional block attention module[C]. European Conference on Computer Vision. Cham: Springer, 2018: 3-19.

[20] Huang L H, Zhao X, Huang K Q. GOT-10 k: A large high-diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5): 1562-1577.

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

[22] Danelljan M, Bhat G, Khan F S, et al. ECO: efficient convolution operators for tracking[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017: 6931-6939.

[23] Danelljan M, Häger G, Khan F S, et al. Learning spatially regularized correlation filters for visual tracking[C]. Proceedings of the IEEE International Conference on Computer Vision. Santiago, 2016: 4310-4318.

[24] Danelljan M, Häger G, Khan F S, et al. Discriminative scale space tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(8): 1561-1575.

[25] Mueller M, Smith N, Ghanem B. A benchmark and simulator for UAV tracking[C]. Proceedings of the European Conference on Computer Vision. Cham: Springer, 2016: 445-461.

[26] Zhu Z, Wang Q, Li B, et al. Distractor-aware Siamese networks for visual object tracking[C]. Proceedings of European Conference on Computer Vision. Munich, 2018: 103-119.

[27] Li X, Ma C, Wu B, et al. Target-aware deep tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, 2019: 1369-1378.

[28] Danelljan M, Robinson A, Khan F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]. Proceedings of the European Conference on Computer Vision. Amsterdam, 2016: 472-488.

[29] Li F, Yao Y J, Li P H, et al. Integrating boundary and center correlation filters for visual tracking with aspect ratio variation[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops. Venice, 2018: 2001-2009.

[30] Li F, Tian C, Zuo W M, et al. Learning spatial-temporal regularized correlation filters for visual tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, 2018: 4904-4913.

[31] Wang N, Song Y B, Ma C, et al. Unsupervised deep tracking[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, 2020: 1308-1317.

[32] Kristan M, Matas J, Leonardis A, et al. The visual object tracking VOT2015 challenge results[C]. Proceedings of the IEEE International Conference on Computer Vision Workshop. Santiago, 2016: 564-586.

[33] Pu S, Song Y, Ma C, et al. Deep attentive tracking via reciprocative learning[C]. The 32nd Conference on Neural Information Processing Systems, NeurIPS. Montreal, 2018: 1931-1941.

[34] Vojir T, Noskova J, Matas J. Robust scale-adaptive mean-shift for tracking[J]. Pattern Recognition Letters, 2014, 49(3): 250-258.

[35] Abdelpakey M H, Shehata M S, Mohamed M M. DensSiam: End-to-end densely-Siamese network with self-attention model for object tracking[C]. The 13th International Symposium on Visual Computing. Las Vegas, 2018: 463-473.

作者简介

仲训杲 (1983–), 男, 副教授, 博士, 从事机器视觉、智能机器人与深度学习理论等研究, E-mail: zhongxungao@163.com;

范东嘉 (1991–), 男, 硕士生, 从事计算机视觉的研究, E-mail: 1173551328@qq.com;

仲训昱 (1980–), 男, 副教授, 博士, 从事智能感知、导航定位与决策规划等研究, E-mail: zhongxunyu@xmu.edu.cn;

周承仙 (1981–), 男, 高级实验师, 硕士, 从事光电检测、机器人视觉等研究, E-mail: zchx258@163.com;

赵晶 (1974–), 女, 教授, 博士, 从事机器视觉、人工神经网络等研究, E-mail: 13606009931@139.com;

刘强 (1989–), 男, 博士后, 博士, 从事人工智能、机器人和精神疾病等研究, E-mail: qiang.liu@psych.ox.ac.uk.