智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

优化分类的弱目标孪生网络跟踪研究

姜文涛, 张大鹏

引用本文:
姜文涛,张大鹏. 优化分类的弱目标孪生网络跟踪研究[J]. 智能系统学报, 2023, 18(5): 984–993.
JIANG Wentao, ZHANG Dapeng. Research on weak object tracking based on Siamese network with optimized classification[J]. *CAAI Transactions on Intelligent Systems*, 2023, 18(5): 984–993.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202211043

DOI: 10.11992/tis.202211043
网络出版地址: https://kns.cnki.net/kcms2/detail/23.1538.TP.20230614.1208.002.html

优化分类的弱目标孪生网络跟踪研究

姜文涛[1], 张大鹏[2]
（1. 辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105; 2. 辽宁工程技术大学 研究生院, 辽宁 葫芦岛 125105）

摘　要: 针对传统孪生网络算法对模糊、低分辨率等弱目标跟踪效果不佳的问题, 提出了优化分类预测的孪生网络算法。首先通过引入可变形卷积模块, 提高骨干网络特征提取能力, 其次在分类分支中引入位置信息, 提升算法对于目标的识别能力, 最后使用轻量级的卷积神经网络进行分类预测和边界预测任务, 在规避多尺度测试的同时, 进一步利用了图像的语义信息, 使跟踪结果具有较高的可信度。在 OTB2015、VOT2018 公共数据集上进行的大量实验表明, 本文算法综合表现优于主流同类算法, 对模糊、形变、快速运动等多种复杂场景具有较好的适应性。

关键词: 计算机视觉; 目标跟踪; 弱目标; 可变形卷积; 先验空间分数; 定位质量评分; 特征提取; 卷积神经网络; 孪生网络
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1673−4785(2023)05−0984−10

中文引用格式: 姜文涛, 张大鹏. 优化分类的弱目标孪生网络跟踪研究 [J]. 智能系统学报, 2023, 18(5): 984–993.
英文引用格式: JIANG Wentao, ZHANG Dapeng. Research on weak object tracking based on Siamese network with optimized classification[J]. CAAI transactions on intelligent systems, 2023, 18(5): 984–993.

Research on weak object tracking based on Siamese network with optimized classification

JIANG Wentao[1], ZHANG Dapeng[2]
(1. College of Graduate School, Liaoning Technical University, Huludao 125105, China; 2. Graduate School, Liaoning Technical University, Huludao 125105, China)

Abstract: To address the problem that traditional Siamese networks are poor in tracking weak objects in blurry and low resolution, this study proposed a Siamese network with optimized classification prediction. First, the feature extraction ability of the backbone network was improved by introducing a deformable convolution module. Second, this algorithm enhances the ability of the backbone network to extract features by introducing the location information in the classification branch. Finally, a lightweight convolutional neural network was used for the prediction of classification and boundary to further utilize the semantic information of the images while avoiding multiscale testing, making the tracking results more reliable. Many experiments have analyzed OTB2015 and VOT2018 datasets, and the results show that the comprehensive performance of this algorithm is better than those of the mainstream similar algorithms, demonstrating excellent adaptability to complex scenes such as motion blur, deformation, and fast motion.

Keywords: computer vision; object tracking; weak object; deformable convolution; prior spatial score; localization quality score; feature extraction; convolutional neural network; siamese network

目标跟踪要求在连续视频序列中, 利用首帧提供的有限信息, 在后续帧中定位目标, 结果一般使用与坐标轴平行的跟踪框表示。目标跟踪技术在自动驾驶[1-2]、增强现实[3]、行人检测[4] 和实时监控[5] 等领域都有良好的应用潜力。在实际应用中目标跟踪也面临着很多由复杂情况带来的挑战, 例如光照条件变化、低分辨率和图片模糊等造成的弱目标情况。因此, 实现在复杂场景下的鲁棒性跟踪仍是计算机视觉领域中一个具有挑战性的任务[6-7]。

随着深度学习的发展, 基于孪生网络的跟踪器逐渐成为主流, 这类算法一般包含两个输入分支, 并使用互相关操作连接这两个输入, 生成响应图, 通过响应图来判断目标在搜索区域的大致

收稿日期: 2022−11−30.　网络出版日期: 2023−06−15.
基金项目: 国家自然科学基金项目(61172144); 辽宁省自然科学基金项目(20170540426); 辽宁省教育厅基金项目(LJYL049).
通信作者: 姜文涛. E-mail: lntuwulue@163.com.

位置。经典算法有 Bertinatto 等[8] 提出的全卷积孪生神经网络(fully-convolutional Siamese network, SiamFC)算法和 Li 等[9] 提出的孪生区域提议(siamese region proposal network, SiamRPN)算法等。SiamFC 算法通过对目标和搜索区域的响应图进行插值, 并在 5 个尺度上对响应图峰值处进行测试来确定目标的具体位置和大小。刘如浩等[10] 在 SiamFC 的基础上引入可变形卷积 (deformable convolution, DC)[11] 提出了 DCSiam 算法, 通过 DC 模块提升单个卷积核的感受野, 进而使骨干网络的特征提取能力得到增强, 同时引入模板更新策略, 在目标特征发生变化时, 依然能够实现有效匹配。但该算法继承了多尺度测试的边界预测策略, 对运动模糊等弱目标场景适应性较差。SiamRPN 算法将区域推荐网络 (region proposal network, RPN) 引入孪生网络, 通过在目标特征图上生成不同尺寸和比例的锚框, 基于锚框预测目标边界, 提升了算法精度。尚欣茹等[12] 在 SiamRPN 中引入导向锚框网络 (guided anchor RPN, GA-RPN), 提出了 SiamGA-RPN算法, 利用深度特征的语义信息指导锚框的生成。这样生成的锚框对目标覆盖更准确, 提高了算法的鲁棒性。但该算法未能充分利用图片语义信息进行分类预测, 在低分辨率、模糊等场景下无法生成高质量锚框。

此外, 上述算法都存在一个共同的问题, 即以标识目标的标注框作为学习目标。标注框通常是平行于坐标轴的矩形框, 为了完整包含目标, 标注框往往比目标主体要大, 因此会包含很多的背景信息, 导致算法在训练过程中会受到干扰, 降低学习效率。基于上述问题, 本文设计了基于分类和回归的一阶段跟踪算法。1)使用孪生全卷积神经网络作为基本框架, 通过引入可变形卷积, 提高骨干网络卷积核的有效感受野, 进而提升骨干网络的特征提取能力。2)在分类训练中引入定位质量信息对训练目标进行加权, 提高真实目标对网络的贡献, 增强模型识别目标的能力。3)在目标预测阶段引入轻量级的卷积神经网络, 充分利用图片语义信息, 并且采用双分支共同预测的网络架构, 将定位和边界回归任务分配到独立的两分支, 使网络具有较高的泛化性。本文算法在 OTB2015、VOT2018 和 LaSOT 等跟踪数据集上都表现出了优秀的跟踪性能。

1　优化分类预测的孪生网络

1.1　整体框架

本文算法整体框架如图 1 所示, 包含特征提取子网络和目标预测子网络, 其中, 模板帧和搜索帧通过互相关操作进行融合。在目标预测子网络中, 使用分类分支和回归分支分别进行位置预测和边界预测, 最终产生目标预测框。

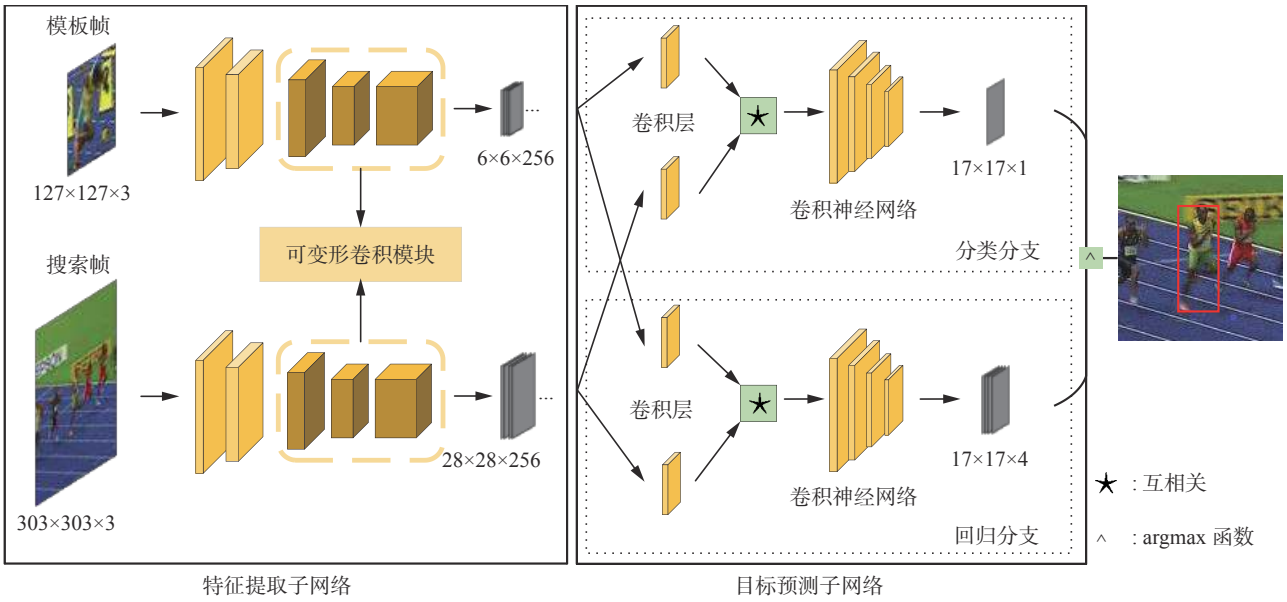


图 1　本文算法框架

Fig. 1　Frame of our algorithm

1.2　可变形卷积

SiamFC 和 SiamRPN 使用 AlexNet 作为骨干网络, 其较浅的层数限制了网络的特征提取能力, 然而使用深层的骨干网络会导致算法的复杂度提高, 影响算法的实时性。本文通过引入可变形卷积模块对 AlexNet 进行改进, 提升骨干网络的特征提取能力的同时, 保证算法的实时性。

卷积运算包含两个步骤: 1)基于规则网格

$\mathcal{R}$在搜索帧x上进行采样；2)将采样值求和并与模板帧w进行卷积。在 AlexNet 中，$\mathcal{R}$定义了一个膨胀为 1，填充为 0 的$m\times m$内核：

$$\mathcal{R}=\{(0,0),(0,1),\cdots,(m-1,m),(m,m)\} \tag{1}$$

输出响应图的每个位置p_0计算方式为

$$y(a)=\sum_{p_n\in\mathcal{R}}w(p_n)\cdot x(p_0+p_n) \tag{2}$$

式中p_n是$\mathcal{R}$上所有位置的枚举。在可变形卷积中，通过添加自适应偏移量卷积层，生成卷积核在不同方向上的偏移Δp_n，并利用偏移集$\{\Delta p_n|n=1,2,\cdots,N\},N=|\mathcal{R}|$对$\mathcal{R}$进行扩充，此时，输出响应图上的每个位置$p_0$的计算方式为

$$y(a)=\sum_{p_n\in\mathcal{R}}w(p_n)\cdot x(p_0+p_n+\Delta p_n) \tag{3}$$

式中：偏移Δp_n通常为小数，因此式(3)可通过双线性插值实现：

$$x(p)=\sum_q G(q,p)\cdot x(q) \tag{4}$$

式中：p表示式(3)中任意位置$p_0+p_n+\Delta p_n$；q是搜索帧x上所有整数空间位置的枚举；$G(\cdot,\cdot)$是二维的双线性插值核，可以被分解为两个一维核：

$$G(q,p)=g(q_x,p_x)\cdot g(q_y,p_y) \tag{5}$$

其中$g(a,b)=\max(0,1-|a-b|)$。

本文可变形卷积使用3×3卷积核，输出偏移图尺寸分别为 (18×34×34)、(18×32×32) 和 (18×30×30) 像素。考虑到目标跟踪任务中，骨干网络的感受野对跟踪效果有重要的影响，感受野过小则无法有效建立语义特征，而过大的感受野会使模型丧失对目标旋转等姿态变化的敏感性。经过仔细分析，确定骨干网络后三层的卷积核尺寸均为 (3×3) 像素，模板帧输出特征图尺寸分别为 (384×10×10) 像素、(384×8×8) 像素和 (256×6×6) 像素，搜索帧输出特征图尺寸分别为 (384×32×32)、(384×30×30) 和 (256×28×28) 像素。如图 2 所示，可变形卷积模块通过引入自适应偏移量，使卷积核能够适应目标的外观变化，有效扩展了卷积核的有效感受野，提高了骨干网络的特征提取能力。

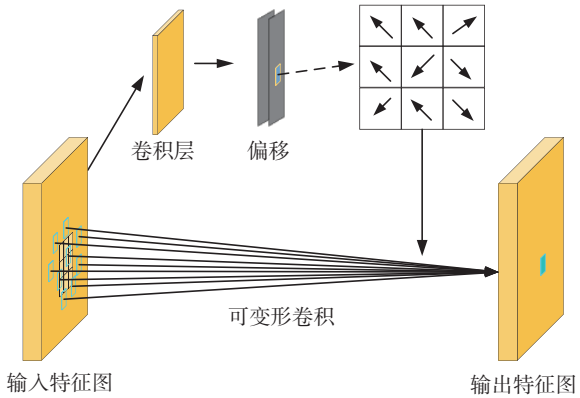


图 2　可变形卷积模块

Fig. 2　Deformable convolution module

1.3　分类预测

由于真实目标种类和姿态变化，其形状不规则，导致标注框中包含有大量背景信息。在预测过程中，这些远离目标中心的区块会产生许多低质量的预测框，导致模型描述目标的能力下降，最终降低跟踪精度和鲁棒性[13]。为了减少背景信息的干扰，提高分类质量，本文在分类训练阶段使用先验空间分数 (prior spatial score, PSS) 作为标注框不同区块的定位质量评分，强化模型的分类能力，其计算方式为

$$S'=\sqrt{\frac{\min(l^*,r^*)}{\max(l^*,r^*)}}\times\sqrt{\frac{\min(t^*,b^*)}{\max(t^*,b^*)}} \tag{6}$$

式中：l^*、t^*、r^*、b^*分别表示正样本区块中心点到标注框左、上、右、下 4 条边的偏移。

如图 3 所示，最终区块的定位质量评分依赖于该区块距离标注框中心的距离，越靠近中心的区块包含目标信息的几率越高，其产生的定位预测分数可信度越高，定位质量评分也越高，而远离中心的区块包含背景信息的几率变高，其对应的定位预测分数可信度越低，相应的定位质量评分会变低，强化了高质量预测区块的贡献。尽管在某些情况下，PSS 分数会导致边缘部分目标区域被降权，但此情况多出现于目标姿态极端不规则的场景，如图 3 中的鸟类头部。在此场景下，标注框边缘的空白区块往往会同步明显增加，稀释该部分目标在训练过程中的作用，因此 PSS 分数在该场景下仍具有较强的指导意义。

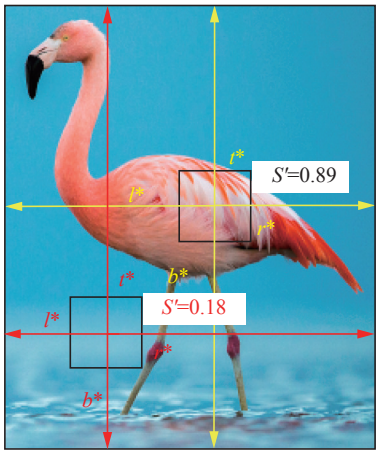


图 3　先验空间分数

Fig. 3　Prior space score

如图 4 所示，现有跟踪算法在训练过程中，直接使用分类标签，边缘的背景区块与中心的目标区块具有相同的贡献，会产生许多不可信的高置

信度分数, 降低模型性能。本文通过引入 PSS 分数对训练目标进行定位强化, 降低边缘低质量区块对模型的影响, 使模型能够聚焦于目标本身, 强化模型的分类能力。

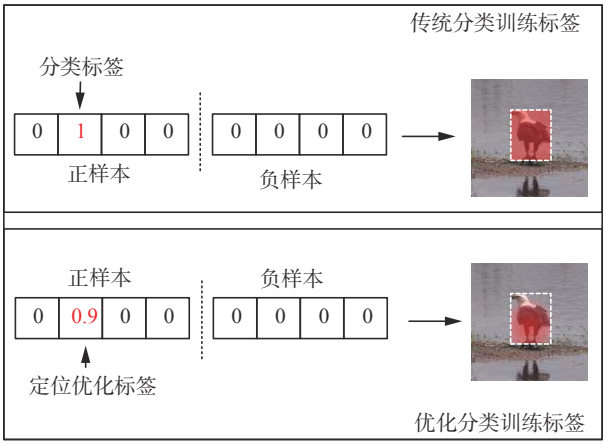


图 4　定位质量优化标签

Fig. 4　Localization quality optimized classification label

在定位强化标签中, 类别标签从离散的$y\in\{0,1\}$变成了连续的$y\in[0,1]$, 因此不能再使用 Focal Loss[14] 作为损失函数。为了支持连续类别标签, 本文使用 QFocal Loss[15] 作为损失函数, 其计算方式为

$$\mathcal{L}_{Q(p_r)}=-\alpha_t|y-p_r|^{\gamma}\cdot((1-y)\ln(1-p_r)+y\ln(p_r)) \quad (7)$$

式中: $y\in[0,1]$表示样本类别, 其中$y=0$表示负样本, $y\in(0,1]$表示正样本。$p_r\in[0,1]$表示预测值, $|y-p_r|^{\gamma}$是动态调节因子, γ的作用是使动态调节过程更加稳定。α_t是正负样本平衡因子, 其计算方式为

$$\alpha_t=\begin{cases}\alpha, y=1\\ 1-\alpha, y=0\end{cases} \quad (8)$$

QFocal Loss 通过对 Focal Loss 进行扩展, 实现了对连续分类标签的支持, 并加入正负样本平衡因子和动态调节因子, 在训练过程中能够使正负训练样本对网络的贡献更加均衡, 同时提高难分类样本对网络的贡献, 增强跟踪器的判别能力。本文α和γ分别设置为 0.25 和 2。

1.4　目标预测子网络

在目标预测子网络中, 目标预测分为分类定位和边界预测两个过程, 分别设计了分类分支和回归分支。前者通过分类任务得到高鲁棒性的相似性评分, 对目标进行定位, 增强跟踪器的判别能力, 后者在前者的基础上, 通过边界框的回归任务对目标的边界进行精确预测, 增强跟踪的精确性。

为了提高骨干网络输出的特征图中语义信息的利用率, 在两个分支中添加了深度为 4 层的卷积神经网络作为头部, 头部网络结构如图 5 所示。最终产生尺度为 (17×17) 像素的特征图。

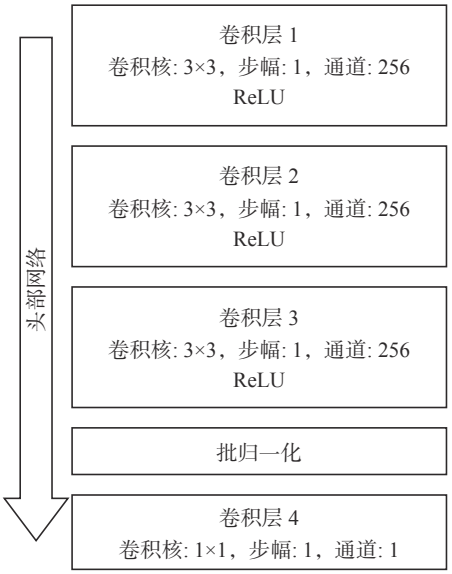


图 5　头部卷积网络结构

Fig. 5　Structure of head convolutional network

其中分类分支将特征图上的每一个点在原图上对应的图像块分类为一个正样本块或负样本块。具体地, 特征图上的每个点(x,y)都可以映射回原图, 对应着以点$\left(\left\lfloor\frac{s}{2}\right\rfloor+xs,\left\lfloor\frac{s}{2}\right\rfloor+ys\right)$为中心的图像块, 其中$s$表示骨干网络的总步幅, 本文中$s=8$。在训练过程中, 如果$(x,y)$对应图像块的中心点在标注框内, 就视作正样本, 否则为负样本。

回归分支与分类分支平行, 基于每个分类结果的正样本点, 预测正样本区块中心相对于标注框的偏移。将正样本点记作$(x_{\text{pos}},y_{\text{pos}})$, 标注框使用左上和右下两点坐标表示, 分别记作$(x_0,y_0)$和$(x_1,y_1)$, 则回归分支的训练标签可以用一个四维向量$\boldsymbol{v}=(l^*,t^*,r^*,b^*)$表示:

$$l^*=\left(\left\lfloor\frac{s}{2}\right\rfloor+sx_{\text{pos}}\right)-x_0,\quad t^*=\left(\left\lfloor\frac{s}{2}\right\rfloor+sy_{\text{pos}}\right)-y_0$$
$$r^*=x_1-\left(\left\lfloor\frac{s}{2}\right\rfloor+sx_{\text{pos}}\right),\quad b^*=y_1-\left(\left\lfloor\frac{s}{2}\right\rfloor+sy_{\text{pos}}\right)$$

回归分支使用 GIoU Loss[16] 作为损失函数, 其计算方式为

$$\mathcal{L}_G=1-I^*_{\text{ou}}+\frac{C-U_{\text{nion}}(B,B^*)}{C} \quad (9)$$

式中: B和B^*分别表示预测框和标注框; C表示预测框和标注框最小外接矩形的面积; $U_{\text{nion}}(\cdot)$表示两个矩形框的并集; I^*_{ou}表示预测框和标注框的交并比。GIoU Loss 在 IoU Loss[17] 的基础上对与标注框没有交集的预测框进行惩罚, 使预测框与标注框不重合时, 回归分支也能继续收敛, 提高了训练速度。

2　实验结果分析

2.1　实验环境及参数配置

本文算法基于开源深度学习框架 Pytorch 实

现，使用 Python3.8 作为开发语言，实验环境为 Manjaro 发行版操作系统。硬件设备环境为 Intel Core i7-10 700@2.7 GHz 八核心处理器，32 GB 内存，使用 NVIDIA Tesla V100 16 GB 显卡加速运算。

采用随机梯度下降法优化模型，使用 ILSVRC-VID/DET，GOT-10k 和 YoutubeBB 数据集作为基础训练数据集进行离线训练。梯度下降动量设置为 0.9，权重衰减为5×10^{-4}，小批次图片数量设置为 32 对，共训练 20 轮，学习率在前 5 轮训练中从10^{-7}线性增长到2×10^{-2}作为预热，并在余下 15 轮中采用余弦退火的方法降低到10^{-7}。

2.2　消融实验

本文算法在 SiamFC 的基础上进行了骨干网络改进、引入定位强化标签和算法结构优化 3 个方面的改进，为了确定改进措施的有效性，以 VOT2018 数据集为基准进行了消融实验。结果如表 1 所示，本文通过添加 DC 模块、添加卷积网络头部、使用定位质量增强标签和添加回归分支等方式，逐步进行实验，通过与上一行的结果比较，分别验证不同改进措施的贡献。表 1 中 A 表示精确度（Accuracy），R 表示鲁棒性（Robustness），分别是 VOT2018 数据集用来评价算法在跟踪过程中对目标的覆盖情况和跟踪稳定性的指标。其中 A 越高越好，R 越低越好。加粗字体表示各列最优结果。

表 1　消融实验结果

Table 1　Results of ablation study

编号	DC模块	卷积网络头部	定位质量标签	回归分支	$A\uparrow$	$R\downarrow$
1	×	×	×	×	0.515	0.573
2	√	×	×	×	0.529	0.501
3	√	√	×	×	0.535	0.428
4	√	√	√	×	0.554	0.253
5	√	√	√	√	**0.586**	**0.206**

编号 2 的实验中，通过添加 DC 模块对骨干网络进行改进。对比编号 1 和编号 2 的实验结果，可以发现引入 DC 模块后，算法精度和鲁棒性分别提高了 0.014 和 0.072，均有较大提升。这是因为 AlexNet 本身特征提取能力较弱，通过应用 DC 模块，不仅增大了卷积核有效感受野，也在一定程度上规避了矩形卷积核无法适应目标形变的缺陷。

编号 3 的实验中，在分类分支中添加卷积网络头部，通过对比编号 2 和编号 3 的实验结果，可以发现算法鲁棒性得到较大提升，达到了 0.073，而精度提升仅有 0.006。这是因为使用卷积网络头部对特征图中的语义信息具有更好的利用率，使算法抵抗噪声干扰的能力更强，跟踪更稳定，但由于依然采用尺度金字塔进行多尺度预测，因此精度提升较为有限。

编号 4 的实验中，引入定位强化标签，通过对比编号 3 和编号 4 的实验，可以发现定位强化标签使算法鲁棒性提升了 0.175，这是由于通过引入定位强化标签，算法能够聚焦于目标本身，因此具有更强的识别能力，同时算法精度提升了 0.019。

编号 5 的实验中，通过添加边界回归分支完善网络结构。新的回归分支避免了多尺度预测，对目标外形轮廓变化具有更好的适应性，使精度提升了 0.032，鲁棒性提升了 0.047。

消融实验的结果证明本文所采取的结构优化、定位强化标签和引入卷积网络头部等措施均对算法性能具有较大提升。

2.3　对比实验及结果分析

2.3.1　弱目标场景实验结果

为了验证本文算法在弱目标场景中的有效性和适应性，从测试数据集中选取弱目标场景的视频序列进行跟踪实验，并选择同类型的 SiamFC、SiamRPN 和基于目标预测准则的鲁棒和精确跟踪（towards robust and accurate visual tracking with target estimation guidelines, SiamFC++）算法[18] 进行定量对比。表 2 统计了各个视频序列对应的场景和具有的挑战属性。

表 2　各序列属性

Table 2　Attributes of video sequences

视频序列	弱目标场景	挑战属性
Blurowl	模糊	SV、MB、FM、IPR
Diving	姿态变化	SV、DEF、IPR
Dragonbaby	模糊、旋转	DV、OCC、MB、FM、IPR、OPR、OV
Ironman	位移、光线变化	IV、SV、OCC、MB、FM、IPR、OV、BC、LR
Jump	姿态变化	SV、OCC、DEF、MB、FM、IPR、OPR
Matrix	弱光，模糊	IV、SV、OCC、FM、IPR、OPR、BC
Redteam	低分辨率	SV、OCC、IPR、OPR、LR
Singer2	复杂光线	IV、DEF、IPR、OPR、BC

特殊序列跟踪结果如表 3 所示，在特殊场景序列中，本文算法成功率和精确率分别达到了 0.655 和 0.845，相比于 SiamFC 分别提高了 31.9% 和 39.9%，综合表现最优。加粗字体表示各行最优结果。

表 3　弱目标场景序列实验结果

Table 3　Experimental results of weak object secne sequences

属性	本文方法	SiamFC	SiamRPN	SiamFC++
成功率	**0.655**	0.336	0.536	0.590
精确率	**0.845**	0.446	0.740	0.785

图 6 给出了不同算法在光线环境复杂、雨天且背景杂乱以及水下场景的跟踪表现。序列 Ironman 中, 由于爆炸等因素, 光照条件变化较为剧烈, 且目标随着姿态变化发生了较大幅度位移和运动模糊, 具有较高的跟踪难度。SiamFC 算法通过对响应图进行插值和尺度金字塔预测目标位置, 在该场景中受背景信息干扰较大, 全程数次丢失目标。SiamRPN 得益于 RPN 网络对目标的二次预测, 在跟踪成功时往往具有较高的覆盖率, 但由于其骨干网络特征提取能力较弱, 同样易受背景干扰, 在第 100 帧以后无法正常跟踪。SiamFC++由于不具备较强的目标分类能力, 因此也在目标同时出现位移、模糊、旋转和光照变化时丢失目标, 并且无法重新定位目标。本文算法由于优化了骨干网络, 并且采用定位强化标签优化分类训练, 具有较强的目标识别能力, 因此能够在复杂背景下完成稳定的跟踪。

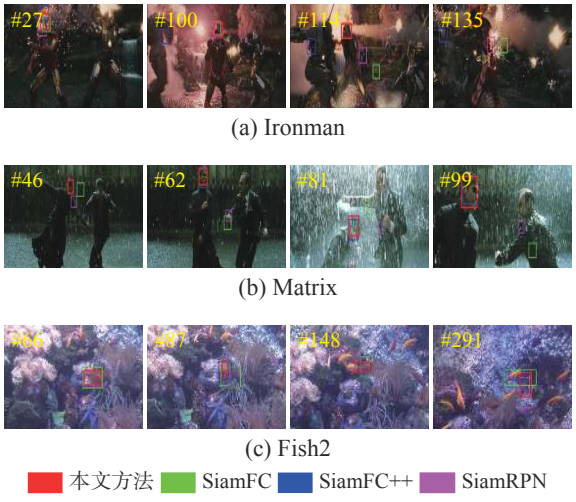


图 6　不同算法在弱目标场景中的跟踪结果

Fig. 6　Tracking results of various algorithms in weak object scenes

Matrix 序列中, 由于暴雨天气和目标运动, 形成了模糊、光照变化弱目标场景。本文算法利用定位强化标签, 提高了对目标的识别能力, 同时使用卷积网络头部增强了算法的分类能力, 能够抵抗各种复杂情况, 较好地跟踪目标。SiamFC 和 SiamRPN 受限于特征提取能力, 对目标识别能力不足, 在第 42 帧、第 62 帧以及第 81~99 帧均发生跟踪失败的情况。SiamFC++同样由于目标识别能力较弱, 虽然能够成功跟踪, 但预测框对目标的覆盖情况不如本文算法。

Fish2 序列是一段水下视频, 背景色调鲜艳且与目标颜色相似, 使目标轮廓不明显, 同时该场景中目标跟踪不仅受到目标本身变化的影响, 还要克服折射、相似目标遮挡等问题, 是非常典型的弱目标场景。可以看到在前 66 帧目标尚未较大变化, SiamFC 预测框就受背景干扰产生偏移, 且在后续跟踪中多次出现相同问题, 这是因为 SiamFC 依赖较浅的骨干网络识别目标, 无法在相似背景下准确分辨目标。SiamFC++在相似目标遮挡时, 也出现了跟踪漂移的问题。SiamRPN 倾向于选择与目标具有较高重叠率的锚框进行拟合, 因此在发生遮挡时容易受锚框诱导, 发生跟踪漂移。本文算法则因为具备较高的识别能力而可以实现较为稳定的跟踪, 且采用了独立的回归分支, 预测框能够较好地适应目标外观和轮廓变化。

2.3.2　OTB2015 数据集实验结果

OTB2015 数据集是目标跟踪领域被广泛采用的数据集, 通过精心选择不同类别和场景的视频片段进行标注, 提供了光照变化(illumination variation, IV)、尺度变化(scale variation, SV)、遮挡(occlusion, OCC)、形变(deformation, DEF)、运动模糊(motion Blur, MB)、快速运动(fast motion, FM)、平面内旋转(in-plane rotation, IPR)、平面外旋转(out-of-plane rotation, OPR)、离开视野(out-of-view, OV)、背景模糊(background-clutter, BC)、低分辨率 (low-resolution, LR)等 11 种对于目标跟踪具有挑战性的属性, 能够较为全面地评估跟踪器性能。采用一次通过评价的方式计算精确率和成功率, 可以作为跟踪器性能对比的公平对比平台。在 OTB2015 数据集上与基于卷积特征的相关滤波跟踪(convolutional features for correlation filter based visual tracking, DeepSRDCF)算法[19]、最大重叠区域的精确跟踪(accurate tracking by overlap maximization, ATOM)算法[20]、SiamFC、SiamRPN、干扰感知的孪生区域提议(distractor-aware siamese network, DaSiamRPN)算法[21]、SiamFC++和基于更深和更宽孪生网络的实时目标跟踪(deeper and wider siamese networks for real-time visual tracking, SiamDWfc)算法[22]进行对比。

对比结果如表 4 所示, 加粗和下划线字体分别表示各列最优和次优结果。本文算法在 OTB2015 上的跟踪成功率为 0.690, 精确率为 0.884, 对比

SiamFC 的成功率(0.586)和精确率(0.772)均有较大提升。对比引入 RPN 网络的 SiamRPN 和 DaSiamRPN 算法，成功率分别提升了 5.8% 和 3.2%，精确率分别提升了 3.4% 和 0.3%。对比迭代优化边界框的 ATOM 算法分别提升了 2.2% 和 1.0%。对比同样使用 AlexNet 作为骨干网络的 SiamFC++ 分别提升了 1.5% 和 0.7%。对比 DeepSRDCF 分别提升了 5.5% 和 2.7%。实验结果表明，本文算法相比于 SiamRPN、SiamFC++、ATOM 等主流算法表现更优。

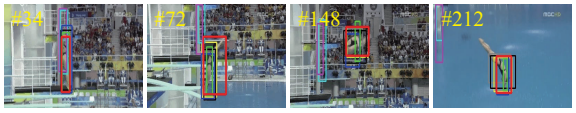
表 4　OTB2015 数据集对比实验结果

Table 4　Comparative experimental results on the OTB2015 dataset

跟踪方法	成功率	精确率	跟踪速度/(f/s)
本文方法	**0.690**	**0.884**	134.319
DeepSRDCF	0.635	0.851	5.383
ATOM	0.668	0.874	25.599
SiamFC	0.586	0.772	73.545
SiamRPN	0.632	0.850	131.330
DaSiamRPN	0.658	0.881	82.647
SiamFC++	0.675	0.877	**145.561**
SiamDWfc	0.627	0.828	66.450

图 7 给出了不同算法在 OTB2015 数据集上部分具有较大挑战性的序列的跟踪表现。Biker 序列中，目标分辨率较低，且具有快速运动、尺度变化和运动模糊等弱目标场景，从第 67~70 帧，目标跳跃过程中发生平面外旋转，DaSiamRPN 受到错误锚框的诱导，丢失目标。DeepSRDCF 无法准确定位目标，发生漂移。在第 70~128 帧，目标在空中旋转并落地，期间快速运动，并伴随着旋转和运动模糊，ATOM、SiamRPN和 SiamFC++均无法正常跟踪目标，DeepSRDCF 算法只能跟踪到部分目标，本文算法通过引入 DC 模块和定位强化标签优化分类能力，对目标识别能力更强，能够在目标快速运动和模糊时稳定识别并定位目标，因此能全程稳定跟踪。

(a) Biker



(b) Diving

(c) Dragonbaby

(d) Girl2

图 7　不同算法在 OTB2015 序列上的跟踪结果

Fig. 7　Tracking results of various algorithms on OTB2015

在 Diving 序列中，主要挑战是目标旋转。在第 34 帧，目标在弹跳阶段发生快速运动，SiamFC 和 SiamDWfc 发生跟踪失败，在第 72 帧，目标准备起跳，身体姿态发生明显变化，SiamRPN 和 DaSiamRPN 由于无法匹配到高质量的锚框，预测框只能覆盖部分目标。在第 72~212 帧中，目标在跳水过程中发生连续旋转和尺度变化，DeepSRDCF 无法适应目标外观变化，SiamRPN 和 DaSiamRPN 由于需要匹配锚框，其预测框变化往往滞后于目标变化，跟踪效果不佳。本文算法由于采用了独立的回归分支，不需要进行多尺度预测，也不受锚框诱导，对目标外观和尺度变化具有较好的鲁棒性，因此预测框能较为完整的覆盖目标，跟踪效果较好。

在 Dragonbaby 序列中，目标在运动过程中发生旋转、遮挡和离开视野等情况，在第 17 帧时，目标颈部的光照变化对跟踪造成干扰，导致 SiamFC、DeepSRDCF 和 SiamFC 发生漂移。在第 44~46 帧，目标快速运动，并伴随平面外旋转和轻微模糊，多种因素影响之下，除本文算法以外所有算法均不能成功跟踪目标，同样的情况在第 84 帧再次出现，本文算法得益于更强的目标识别能力，可以稳定识别目标，实现全程跟踪。

在 Girl2 序列中，当目标在第 106~128 帧期间被行人完全遮挡，目标恢复后，本文算法最先重新识别到目标，SiamRPN 和 DaSiamRPN 则选择了错误的锚框进行拟合，跟踪失败，SiamFC 和 SiamDWfc 同样跟踪到了错误的目标。在第 666 帧时，目标与行人重合，并且背景杂乱，SiamFC++ 预测框偏移到了干扰目标上，SiamFC 跟踪失败。在第 1 406 帧时，目标在运动中发生尺度变化，本文算法能够稳定识别目标，且采用独立回归分支，能较好适应外观变化，因此跟踪效果较好，其

他算法均受到不同程度干扰, 不能很好地跟踪目标。

2.3.3　VOT2018 数据集实验结果

VOT2018 数据集包含 60 个视频序列, 采用四点标注法标注目标, 更贴近直觉, 同时难度也比 OTB2015 更高, 更能体现不同算法之间的性能差距。选择 DeepSRDCF、SiamFC、双重孪生网络实时目标跟踪(a twofold siamese network for real-time object tracking, SA_Siam)算法[23]、SiamRPN、DaSiamRPN、ATOM 和 SiamFC++等主流算法进行对比。

对比实验结果如表 5 所示, 加粗和下划线字体分别表示各列最优和次优结果。本文算法的平均期望重叠率(expected average overlap, EAO)达到了 0.413, 在近些年主流同类算法和非同类算法中综合表现最优。对比实验数据可知, ATOM 算法准确性较高, 这是因为 ATOM 算法在预测目标边界时, 会生成多个预测框进行迭代优化, 因此可以获得准确性较高的预测框, 使跟踪准确性较高, 但会导致算法实时性下降, 跟踪速度仅有 25.83 帧/秒。本文采用独立回归分支和卷积网络头部预测目标边界, 对目标形变适应性较强, 因此跟踪准确性达到了 0.586, 与 VOT2018 冠军算法 SiamRPN 持平, 且鲁棒性更好。

表 5　VOT2018 数据集对比实验结果

Table 5　Comparative experimental results on the VOT2018

跟踪方法	EAO ↑	A ↑	R ↓	跟踪速度/f/s
本文方法	**0.413**	0.586	0.206	129.46
DeepSRDCF	0.154	0.492	0.707	5.10
SiamFC	0.188	0.506	0.585	69.58
SA_Siam	0.337	0.566	0.258	40.45
SiamRPN	0.383	0.586	0.276	128.71
DaSiamRPN	0.326	0.569	0.337	78.64
ATOM	0.401	**0.590**	0.204	25.83
SiamFC++	0.400	0.556	**0.183**	**141.35**

图 8 给出了不同算法在 VOT2018 数据集上部分具有较高难度的视频序列上的跟踪结果。序列 Dinosaur 中, 跟踪器需要克服目标快速运动、旋转和模糊等问题, 在第 38 帧, 目标轮廓受到背景干扰变得模糊, SiamFC 不能正确识别目标。在第 121 帧, 目标发生模糊且发生旋转, 本文算法依然能稳定识别并跟踪目标, 而 ATOM 和 SiamFC 算法均跟踪失败。在第 158 帧和 281 帧, 目标出现较为剧烈的变化, 且光照杂乱, 背景干扰较强, SiamRPN 和 DaSiamRPN 采取的拟合锚框策略不能再稳定跟踪目标, 而本文算法具有较强的目标识别能力, 可以正常跟踪目标。

图 8　不同算法在 VOT2018 序列上的跟踪结果

Fig. 8　Tracking results of various algorithm on VOT2018

在 Fernando 序列中, 目标受到严重遮挡, 且伴随强烈的光照和外形变化。在第 55~86 帧, 目标缓慢移动过程中, 各个对比算法均不能较好地适应目标轮廓变化导致预测框只能覆盖部分目标, 本文算法预测框对目标覆盖良好。第 151 帧出现了明显的光照变化, SiamFC、ATOM 和 SA_Siam 将地面和部分背景当作目标, 预测框出现偏移, SiamRPN、DaSiamRPN、DeepSRDCF 和 SiamFC++都只能跟踪到部分目标, 只有本文算法正确识别到猫尾, 体现了本文算法较强的目标识别能力。第 283 帧目标与干扰物分离, 本文算法最先重新跟踪到目标, 并且能够完整覆盖目标, SiamFC 识别到错误目标, 其余算法均只能跟踪到部分目标。

在 Gymnastics3 序列中, 目标全程快速运动, 出现了模糊弱目标, 且背景较为杂乱。在前 22 帧目标在奔跑过程中, 身体轮廓发生改变, 只有本文算法和 SiamFC++能够正确识别目标腿部, 实现较为完整的覆盖, 其余算法均只能识别目标上半身。在第 34~91 帧目标起跳和落地过程中发生旋转和运动模糊, 只有本文算法能够较好地适应目标姿态变化, 其余算法预测框变化均滞后于目标姿态变化, 导致对目标覆盖情况不佳。

在 Motocross1 序列中, 主要挑战是目标的快

速运动、旋转和背景杂乱。DeepSRDCF、SiamFC 和 ATOM 由于对目标特征提取能力较弱，在第 36 帧背景杂乱时不能正确识别目标，跟踪失败。在第 75~94 帧目标持续旋转，且背景不断变化，本文算法依靠较强的目标识别能力能够持续稳定跟踪目标，且预测框能较好覆盖目标，其余算法虽然能成功跟踪，但预测框对目标覆盖情况不够理想。第 147 帧目标尺度发生较大变化，本文算法预测框能够适应目标外观变化，SiamFC++预测框过大，包含较多背景，SiamFC 和 DeepSRDCF 则预测框过小，不能完整包含目标。

2.3.4　LaSOT 数据集实验结果

为了更好验证本文算法的有效性以及在较长序列跟踪时的稳定性，引入了大规模单目标跟踪数据集 LaSOT。LaSOT 数据集对 1 400 个视频序列进行了密集标注，并且平均视频序列长度达到了 2 512 帧，具有较高的难度，能够更好地验证跟踪算法的稳定性和泛化性，采用一次通过评价计算成功率和精确率。选择 SiamDWfc、SiamFC、ATOM、SiamFC++、DaSiamRPN、一种统一的快速目标跟踪和分割方法（fast online object tracking and segmentation: a unifying approach, SiamMask）[24] 和多域卷积神经网络（multi-domain convolutional neural network, MDNet）算法[25] 等在 LaSOT 数据集上具有较好表现的算法进行对比。

对比试验结果如表 6 所示，加粗和下划线字体分别表示各列最优和次优结果。采后多种改进措施后，本文算法在 LaSOT 提供的长序列跟踪场景中达到了 0.550 的成功率和 0.635 的精确率，相比于 DaSiamRPN 分别提高了 3.0% 和 3.5%，相比于基线算法 SiamFC 分别提高了 13.5% 和 21.4%，在近些年的同类型算法中成功率和精确率均表现最优，同时算法实时性也表现较好。

表 6　LaSOT 数据集对比实验结果

Table 6　Comparative experimental results on the LaSOT

跟踪方法	成功率	精确率	跟踪速度/f/s
本文方法	**0.550**	**0.635**	<u>130.32</u>
SiamFC	0.336	0.420	70.81
SiamDWfc	0.347	0.437	65.52
ATOM	0.499	0.570	26.47
DaSiamRPN	<u>0.515</u>	<u>0.605</u>	79.70
SiamFC++	0.500	0.571	**143.34**
SiamMask	0.467	0.552	29.93
MDNet	0.394	0.460	1.65

3　结束语

针对孪生网络跟踪算法对模糊和低分辨率等情况下的弱目标跟踪效果不佳的问题，以优化模型分类能力的思路，从特征提取和语义特征应用等方面对孪生全卷积网络进行改进。应用可变形卷积模块改进骨干网络，并对网络结构进行优化，引入卷积网络头部和独立回归分支，同时利用定位质量评分对分类训练进行优化，提高算法对目标的识别能力。在 OTB2015、VOT2018 和 LaSOT 公共数据集上的大量对比实验证明所提算法在模糊、低分辨率等弱目标场景中能够实现较为稳定的跟踪，综合表现优于当前同类主流算法，且算法实时性较好。

所提算法依然存在改进空间，由于采取不更新模板的策略，在目标遮挡物消失时能够迅速地重新发现和定位目标，但在遮挡期间有可能受到相似干扰目标的诱导，导致跟踪失败，因此进一步的研究方向是引入合适的模板更新策略，提高算法的抗遮挡干扰能力。

参考文献：

[1] 李家宁, 田永鸿. 神经形态视觉传感器的研究进展及应用综述 [J]. 计算机学报, 2021, 44(6): 1258–1286.
LI Jianing, TIAN Yonghong. Recent advances in neuromorphic vision sensors: a survey[J]. Chinese journal of computers, 2021, 44(6): 1258–1286.

[2] 朱向雷, 王海弛, 尤翰墨, 等. 自动驾驶智能系统测试研究综述 [J]. 软件学报, 2021, 32(7): 2056–2077.
ZHU Xianglei, WANG Haichi, YOU Hanmo, et al. Survey on testing of intelligent systems in autonomous vehicles[J]. Journal of software, 2021, 32(7): 2056–2077.

[3] 韩玉仁, 李铁军, 杨冬. 增强现实中三维跟踪注册技术概述 [J]. 计算机工程与应用, 2019, 55(21): 26–35.
HAN Yuren, LI Tiejun, YANG Dong. Overview of 3D tracking registration technology in augmented reality[J]. Computer engineering and applications, 2019, 55(21): 26–35.

[4] 王梦来, 李想, 陈奇, 等. 基于 CNN 的监控视频事件检测 [J]. 自动化学报, 2016, 42(6): 892–903.
WANG Menglai, LI Xiang, CHEN Qi, et al. Surveillance event detection based on CNN[J]. Acta automatica sinica, 2016, 42(6): 892–903.

[5] TANG Siyu, ANDRILUKA M, ANDRES B, et al. Multiple people tracking by lifted multicut and person re-identification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 3701−3710.

[6] ZHU Zheng, WU Wei, ZOU Wei, et al. End-to-end flow correlation tracking with spatial-temporal attention[C]//

2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 548−557.

[7] 张长弓, 杨海涛, 王晋宇, 等. 基于深度学习的视觉单目标跟踪综述 [J]. 计算机应用研究, 2021, 38(10): 2888–2895.
ZHANG Changgong, YANG Haitao, WANG Jinyu, et al. Survey on visual single object tracking based on deep learning[J]. Application research of computers, 2021, 38(10): 2888–2895.

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

[9] LI Bo, YAN Junjie, WU Wei, 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.

[10] 刘如浩, 张家想, 金辰曦, 等. 基于可变形卷积的孪生网络目标跟踪算法 [J]. 控制与决策, 2022, 37(8): 2049–2055.
LIU Ruhao, ZHANG Jiaxiang, JIN Chenxi, et al. Target tracking based on deformable convolution Siamese network[J]. Control and decision, 2022, 37(8): 2049–2055.

[11] DAI Jifeng, QI Haozhi, XIONG Yuwen, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 764−773.

[12] 尚欣茹, 温尧乐, 奚雪峰, 等. 孪生导向锚框 RPN 网络实时目标跟踪 [J]. 中国图象图形学报, 2021, 26(2): 415–424.
SHANG Xinru, WEN Yaole, XI Xuefeng, et al. Target tracking system based on the Siamese guided anchor region proposal network[J]. Journal of image and graphics, 2021, 26(2): 415–424.

[13] TIAN Zhi, SHEN Chunhua, CHEN Hao, et al. FCOS: fully convolutional one-stage object detection[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2020: 9626−9635.

[14] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 2999−3007.

[15] LI Xiang, WANG Wenhai, WU Lijun, et al. Generalized focal loss: learning qualified and distributed bounding boxes for dense object detection[J]. Advances in neural information processing systems, 2020, 33: 21002–21012.

[16] REZATOFIGHI H, TSOI N, GWAK J, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 658−666.

[17] JIANG Borui, LUO Ruixuan, MAO Jiayuan, et al. Acquisition of localization confidence for accurate object detection[C]//European Conference on Computer Vision. Cham: Springer, 2018: 816−832.

[18] XU Yinda, WANG Zeyu, LI Zuoxin, et al. SiamFC++: towards robust and accurate visual tracking with target estimation guidelines[EB/OL]. (2019−11−14)[2021−01−01]. https://doi.org/10.48550/arXiv.1911.06188.

[19] DANELLJAN M, HÄGER G, KHAN F S, et al. Convolutional features for correlation filter based visual tracking[C]//2015 IEEE International Conference on Computer Vision Workshop. Santiago: IEEE, 2016: 621−629.

[20] DANELLJAN M, BHAT G, KHAN F S, et al. ATOM: accurate tracking by overlap maximization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 4655−4664.

[21] ZHU Zheng, WANG Qiang, LI Bo, et al. Distractor-aware Siamese networks for visual object tracking[C]//European Conference on Computer Vision. Cham: Springer, 2018: 103−119.

[22] ZHANG Zhipeng, PENG Houwen. Deeper and wider Siamese networks for real-time visual tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 4586−4595.

[23] HE Anfeng, LUO Chong, TIAN Xinmei, et al. A twofold Siamese network for real-time object tracking[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4834−4843.

[24] WANG Qiang, ZHANG Li, BERTINETTO L, et al. Fast online object tracking and segmentation: a unifying approach[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2020: 1328−1338.

[25] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4293−4302.

作者简介:

姜文涛, 副教授, 博士, 主要研究方向为图像处理、模式识别、人工智能。参与国家及省级项目 2 项, 发表学术论文 20 余篇。

张大鹏, 硕士研究生, 主要研究方向为图像处理、模式识别、人工智能。