DOI:10.19651/j.cnki.emt.2210932

基于强化特征学习和表达策略的孪生网络跟踪算法*

符 强[1,2,3] 王 阳[1,2,3] 纪元法[1,2,3] 任风华[1,2,3]
(1.桂林电子科技大学广西精密导航技术与应用重点实验室 桂林 541004; 2.桂林电子科技大学信息与通信学院 桂林 541004;
3.卫星导航定位与位置服务国家地方联合工程研究中心 桂林 541004)

摘 要:针对基于全卷积孪生网络跟踪算法在面对相似物干扰、光照变化等复杂环境时容易出现跟踪漂移的问题,本文在分析与实验基础上提出如下特征强化策略。首先,将改良的深度卷积神经网络VGG16引入跟踪框架来提高模型的特征学习能力;其次,针对单一特征无法充分描述目标信息,且对干扰物比较敏感的问题,本文设计一种特征增强模块,由浅至深融合不同层次语义信息来提高特征的表达能力;最后,提出一种轻量级的三元注意力机制,帮助模型自适应关注优势特征,进一步提高了模型在复杂环境下的鲁棒性。将上述策略应用到全卷积孪生网络算法上取得了显著的效果。在OTB100数据集上,本文算法成功率曲线下面积较基准算法提升了15.1%,距离精度提升了16.3%,在复杂环境下也能对目标进行有效跟踪。

关键词:目标跟踪;孪生网络;特征提取;注意力机制
中图分类号:TN391.41 **文献标识码**:A **国家标准学科分类代码**:520.604

Siamese network tracking algorithm based on reinforcement feature learning and expression strategy

Fu Qiang[1,2,3] Wang Yang[1,2,3] Ji Yuanfa[1,2,3] Ren Fenghua[1,2,3]
(1. Guangxi Key Laboratory of Precision Navigation Technology and Application, Guilin University of Electronic Technology, Guilin 541004, China; 2. School of Information and Communication, Guilin University of Electronic Technology, Guilin 541004, China; 3. National & Local Joint Engineering Research Center of Satellite Navigation Positioning and Location Service, Guilin 541004, China)

Abstract: Aiming at the problem that the tracking algorithm based on the fully convolutional siamese network is easy to tracking drift in the face of complex environments such as analog interference and illumination changes, this paper proposes the following strategies to optimize features on the basis of analysis and experiments. First, the deep convolutional neural network VGG16 is introduced into the tracking framework to improve the feature extraction ability of the model. Then, aiming at the problem that a single feature cannot adequately describe the target information and is sensitive to interferences, this paper designs a feature enhancement module, which integrates different levels of semantic information from shallow to deep to improve the expressiveness of features. Finally, a lightweight triple attention is proposed to help the model adaptively focus on dominant features and further improve the robustness of the model in complex environments. Applying the above strategies to the fully convolutional siamese network algorithm has achieved remarkable results. On the OTB100 dataset, compared with benchmark algorithm, the area under the success rate curve of the algorithm in this paper is increased by 15.1%, and the distance accuracy is increased by 16.3%, and the target can also be effectively tracked in complex environment.

Keywords: object tracking; siamese network; feature extraction; attention mechanism

0 引 言

目标跟踪是计算机视觉学科中十分热门的研究方向[1],同时也是一项具有挑战性的研究工作。通过对目标的外观和运动重建,跟踪器可以预测目标的运动轨迹,从而获取目标的位置信息,基于此特性,目标跟踪技术广泛应用

收稿日期:2022-08-02
*基金项目:国家自然科学基金(61561016,61861008)、广西科技厅项目(桂科AA19182007)、"认知无线电与信息处理"教育部重点实验室(CRKL200108)、广西精密导航技术与应用重点实验室(DH201901)、桂林电子科技大学研究生教育创新计划项目(2022YCXS050)资助

在智能交通监控、智能人机交互、军事侦察等领域[2]。随着研究深入,各种优秀算法层出不穷,其中,基于孪生网络的跟踪算法以其兼顾精度与速度的特点吸引了越来越多的研究者。SINT[3]和 SiamFC[4]是较早将孪生网络应用到跟踪任务的,它们将跟踪问题转换成相似性问题,通过学习一个相似度度量函数来求解问题。SiamFC 问世后,一系列基于此的算法也相继被研究出来。SiamRPN[5]借鉴检测任务中 Faster-RCNN[6]算法,将区域提议网络(region proposal network,RPN)引入到跟踪框架,利用 RPN 进行前景与背景分类和边界框回归,有效提高了预测边界框的准确性,同时也避免了 SiamFC 中的多尺度检测,显著提高了跟踪器的速度,但对所有特征使用统一处理方式,没有突出不同特征的优势;DaSiamRPN[7]在离线训练阶段引入一种有效的抽样策略来控制训练数据集的分布,以此来提升训练效果;SiamDW[8]针对特征提取网络问题,提出一种深层网络结构,利用深层网络提取的特征信息更具有辨别性的优势来提高跟踪器性能;Valmadre 等[9]提出的 CFNet 算法,将相关滤波(correlation filter,CF)引入到孪生网络框架中,将 CF 解释为可微的卷积神经网络层,让参数可以通过 CF 反向传递至卷积神经网络特征,提升了算法精度,但在面对遮挡问题时,存在鲁棒性不足的问题。

基于全卷积孪生网络跟踪算法虽然取得了不错的效果,但其较浅的特征提取网络和单一的特征处理方式使模型在面对复杂环境时表现一般。综上,本文基于孪生网络思想提出融合多层次语义和三元注意力机制等特征强化策略来解决上述问题,主要工作如下:1)针对网络特征学习能力不足的问题,使用改良的 VGG16 作为主干网络。同时,在网络中设计裁剪块消除填充操作带来的潜在位置偏差影响,并结合空洞卷积增大感受野,获取更加丰富的上下文信息。2)鉴于不同层次特征对目标的描述角度不同,设计一种由浅至深的语义信息融合模块,提升特征对目标表述能力。3)提出一种轻量级的三元注意力机制,通过关注特征的维度和空间属性,自适应优化特征表达,提高模型在面对复杂环境时的鲁棒性。

1　全卷积孪生网络算法

SiamFC 将跟踪问题转换成相似性问题,通过端到端训练一个模型来求解该问题。该算法使用 AlexNet 作为特征提取网络,采用离线方式训练跟踪模型网络参数。在跟踪阶段,首先对第 1 帧目标提取深度特征,得到模板图像对应的特征,然后对后续帧输入的搜索区域进行特征提取,最后对模板特征和搜索特征进行互相关操作得到响应矩阵。网络框架如图 1 所示,网络由模板分支和搜索分支组成,且两个分支共享权重。算法利用全卷积网络特点,即两个分支的输入可以是不同尺寸,因此可以向网络提供更大的搜索图像作为输入,计算更多子窗口与模板的相似度,相似度量函数 $f(\cdot,\cdot)$ 如式(1)所示。

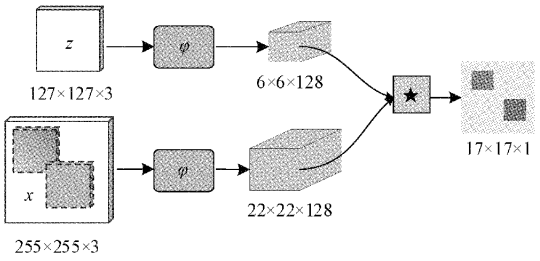


图 1　SiamFC 算法网络框架

$$f(z,x)=\varphi(z)*\varphi(x)+b1 \tag{1}$$

其中,z 是输入模板图像,x 是输入搜索图像,$\varphi(\cdot)$ 是每个分支对应的特征提取网络,$*$ 表示互相关操作,$b1$ 表示偏移量。响应矩阵中得分最高的位置即为目标预测位置。

2　本文算法

基于孪生网络思想,本文采用层数更深的 VGG16[10]作为特征提取主干网络,在网络中嵌入层次语义融合模块和三元注意力机制来优化特征表达,获取丰富的目标描述子。算法的整体流程如图 2 所示。

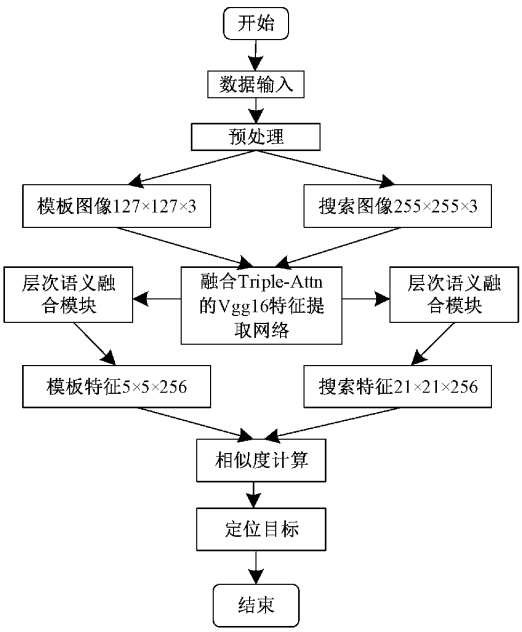


图 2　本文算法流程

2.1　网络整体参数

虽然增加网络深度已被证明有利于提高模型性能,但将具有连续卷积跨步的 VGG16 应用到跟踪任务仍然存在不足:1)网络中的填充操作破坏了网络的平移不变性,带来潜在的位置偏差,对处在边缘位置的目标影响很大,如图 3 所示,整体网络框架如图 4 所示,Z 表示模板图像,A 是搜索图像的特征图,B 是搜索图像中目标移动到边缘位置的特征图。根据式(1)可知:

$$\begin{cases} f(Z,A) = \varphi(Z) * \varphi(A) + b1 \\ f(Z,B) = \varphi(Z) * \varphi(B) + b1 \end{cases} \tag{2}$$

当目标移动到边缘部位时，搜索图像中 B 处方框可能包含填充信息，而模板图像中 Z 处方框只包含目标自身信息，这会导致模板特征和搜索特征之间最终的相似度求解存在误差，导致相同目标移动后在响应图中的得分是不同的，即 $f(Z,A) \neq f(Z,B)$，则映射回原图中的位置会有偏差。

2）在工作[8,11]中的实验表明，网络步幅 4 或 8 性能要优于 16 或 32，且基于孪生网络的跟踪器需要详细的空间

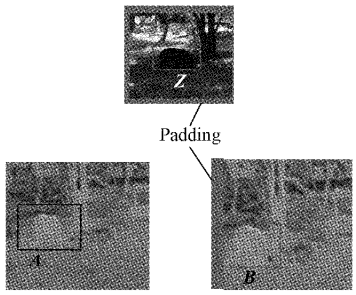


图 3　填充影响示意

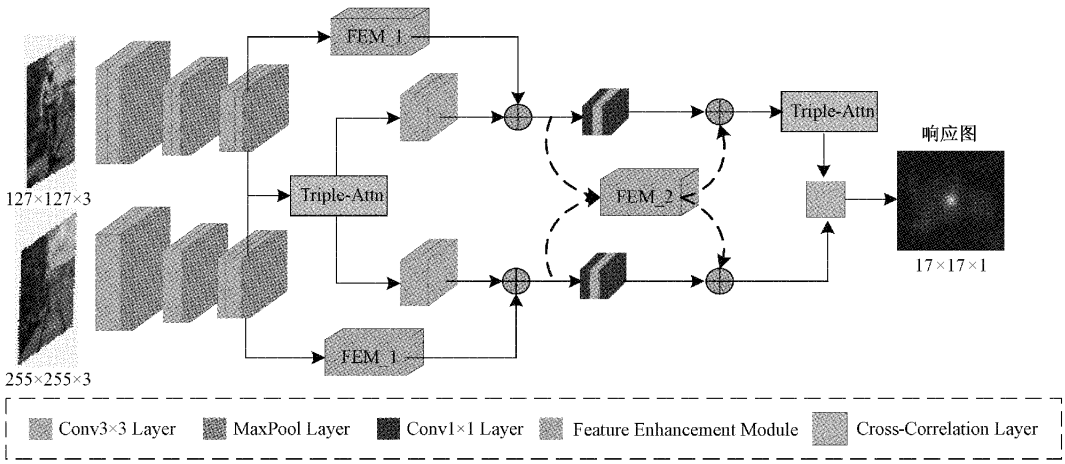


图 4　本文算法网络框架

信息来执行预测任务，原始 VGG16 的网络步幅为 32，网络最后输出的特征空间分辨率太低，不利于定位目标。

基于上述实验分析，本文对 VGG16 改良如下：针对填充干扰问题，借鉴工作[8]，设计裁剪块将受填充操作影响的最外层特征裁剪掉；为了避免网络步幅过长导致输出特征分辨率过小，只保留网络的前 3 层下采样，即将网络步幅缩短至 8。同时，在 Layer5 中结合空洞率（dilation）为 2 的空洞卷积[12]来解决因缩短网络步幅带来感受野不充足的问题，让输出特征的每一个像素位置都包含上一层特征较大范围的信息，以此来捕获多尺度上下文信息。空洞卷积支持感受野的指数级扩展而不会丢失分辨率，每个元素感受野大小计算公式如式（3）所示：

$$F_{i+1} = F_i * 2^i k_i \quad i = 0,1,\cdots,n-2 \tag{3}$$

式中：k 是大小为 3×3 的滑动滤波器，F 表示经过卷积后输出的特征。

由公式可以看出，每一层特征中元素的感受野大小都是基于上一层特征卷积结果得到的。当使用空洞率大小为 2 的幂次方的滑动滤波器在第 F_i 层进行卷积操作后，则 F_{i+1} 层特征中每个元素的感受野大小为 $(2^{i+2}-1)\times(2^{i+2}-1)$。空洞卷积增大感受野效果如图 5 所示，最外层矩形框表示感受野大小。改进的 VGG16 网络参数和各层对应操作如表 1 所示。网络总共分为 5 层，其中 3×3、1×1 分别代表卷积层的核大小，所有卷积层步幅均为 1，crop 表示特征裁剪块；Max 代表核大小为 2，步幅为 2 的最大池化层，Triple-Attn 表示三元注意力机制。

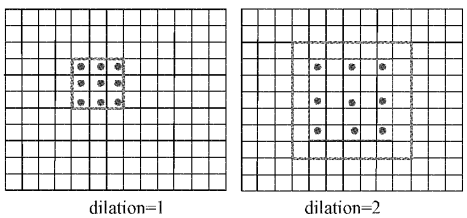


图 5　感受野增大效果

2.2　三元注意力机制（Triple-Attn）

注意力机制因其能为卷积神经网络结构带来性能增益而被广泛应用在计算机视觉任务中，包括分类任务、检测任务、跟踪任务等。注意力机制的本质作用是对获取的信息进行分析，突出其中重要信息，将次要信息作为辅助信息对视觉任务进行分析理解[13]。跟踪任务中常用的注意力机制有 CBAM[14]，ECA-Net[15]，SENet[13] 等，可分为通道注意力机制和空间注意力机制。根据跟踪任务的特性及对特征属性的要求，本文在上述工作基础上提出一种轻量级的三元注意力机制（triple attention mechanism），经过 Triple-Attn 处理的特征能显著提高对特定对象的辨别能力，提升跟踪器在复杂环境下的鲁棒性。

通道注意力关注的是特征图不同通道对不同目标描述子的不同响应程度，由于特征图的每个通道都可以视为

表 1　网络参数及各层对应参数

网络层	操作	模板图像大小	搜索图像大小
—	—	127×127×3	255×255×3
Layer1:	3×3	127×127×64	255×255×64
	crop	125×125×64	253×253×64
	3×3	125×125×64	253×253×64
	crop	123×123×64	251×251×64
	Max	61×61×64	125×125×64
Layer2:	3×3	61×61×128	125×125×128
	crop	59×59×128	123×123×128
	3×3	59×59×128	123×123×128
	crop	57×57×128	121×121×128
	Max	28×28×128	60×60×128
Layer3:	3×3	28×28×256	60×60×256
	crop	26×26×256	58×58×256
	3×3	26×26×256	58×58×256
	crop	24×24×256	56×56×256
	3×3	24×24×256	56×56×256
	crop	22×22×256	54×54×256
	Max	11×11×256	27×27×256
Triple-Attn	—		
Layer4:	3×3	11×11×512	27×27×512
	crop	9×9×512	25×25×512
	3×3	9×9×512	25×25×512
	crop	7×7×512	23×23×512
Layer5:	1×1	7×7×256	23×23×256
	3×3	5×5×512	21×21×512
	1×1	5×5×256	21×21×256
Triple-Attn (模板分支)	—		

特征检测器[16],根据响应程度自适应赋予不同特征通道不同的权重来调整通道对不同目标重视程度,以此来表述输入特征通道中“什么”是对任务有意义的。空间注意力关注的是特征通道内部信息之间的依赖关系,获取特征图中不同空间位置的信息权重,它是对通道注意力的补充,向通道注意力传递“哪里”是有意义的位置信息。本文提出的三元注意力机制将通道注意力和空间注意力各自聚焦点进行有效嵌入,具体实现过程如图 6 所示。

在通道注意力中,为了获取全局的通道描述矩阵,首先对卷积层输出特征 $F \in \mathbb{R}^{H\times W\times C}$ 进行二维全局平均池化,如式(4)所示。

$$\boldsymbol{f}_{gap}=\frac{1}{WH}\sum_{i=1,j=1}^{W,H}F_{i,j} \tag{4}$$

其中,W 和 H 分别是特征图的宽和高,$\boldsymbol{f}_{gap}=(f_1,f_2,\cdots,f_c)$,$f_i\in\boldsymbol{R}$。本文通道注意力没有使用常用的多层感知机方式来预测通道权重,而是引用工作[15]采用相邻通道交互的方法,这种方法不仅可以保证通道与权重之间的直接联系,同时也可以显著降低模型参数量,保证实时性。通道之间的交互可以通过一维卷积简单实现,且交互通道的数量 K 也可以根据输入特征通道数自适应调整,具体计算如式(5)所示。

$$K=\left|\frac{\ln(C)}{\lambda\ln 2}+\frac{b}{\lambda}\right|_{odd} \tag{5}$$

式中:$|x|_{odd}$ 表示距离 x 最近的奇数,C 为输入特征通道数。根据通道数取值的特性,在实验中取 $\lambda=2$,$b=1$。经过 sigmoid 激活函数后得到通道权重 $\boldsymbol{W}=[\omega_1,\omega_2,\cdots,\omega_c]$,最后将权重信息 $\boldsymbol{W}$ 嵌入到 F 中得到通道注意力输出 $\boldsymbol{\alpha}\in\mathbb{R}^{H\times W\times C}$。

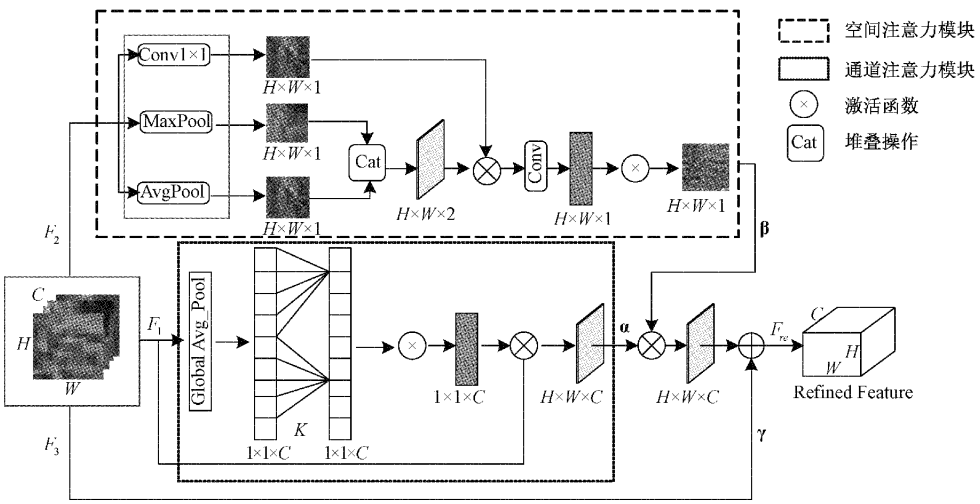


图 6　三元注意力机制设计实现

沿通道维度应用池化操作可以有效凸显信息区域[17]。因此,在空间注意力中,本文将卷积层输出特征 F 沿着通道维度分别进行平均池化和最大池化操作生成二维空间注意力特征图 $\boldsymbol{f}_{avg}\in\mathbb{R}^{H\times W\times 1}$ 和 $\boldsymbol{f}_{\max}\in\mathbb{R}^{H\times W\times 1}$,同时将 F 通过一个 1×1 卷积融合各通道信息生成特征描述符 $\boldsymbol{f}\in\mathbb{R}^{H\times W\times 1}$。将 $\boldsymbol{f}_{avg}$,$\boldsymbol{f}_{\max}$ 堆叠后与 $\boldsymbol{f}$ 相乘再输入核大小为 7×7 的卷积层和 Sigmoid 激活函数得到含有位置信息权重的空间特征 $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}=\begin{pmatrix}\beta_{11} & \cdots & \beta_{1W}\\ \vdots & \ddots & \vdots\\ \beta_{H1} & \cdots & \beta_{HW}\end{pmatrix} \tag{6}$$

其中，β_{ij} 表示空间注意力特征上位置 (i,j) 的信息权重。三元注意力机制总过程如式(7)所示：

$$\begin{cases}\boldsymbol{\alpha}=F_1\otimes\sigma(Conv1D(\boldsymbol{f}_{gap}(F_1)))\\ \boldsymbol{\beta}=\sigma(f_{Conv2d}^{7\times7}(cat[\boldsymbol{f}_{\max},\boldsymbol{f}_{avg}]\otimes f_{Conv2d}^{1\times1}(F_2)))\\ \boldsymbol{\gamma}=F_3\\ F_{re}=\boldsymbol{\alpha}\otimes\boldsymbol{\beta}\oplus\boldsymbol{\gamma}\end{cases} \tag{7}$$

式中：$F_1=F_2=F_3=F\in\mathbb{R}^{H\times W\times C}$，$f_{Conv2d}^{7\times7}$ 和 $f_{Conv2d}^{1\times1}$ 分别表示核大小为 7×7 和 1×1 的卷积层，σ 是激活函数。

三元注意力机制将通道注意力输出 $\boldsymbol{\alpha}$ 作为主干信息，包含位置信息权重的空间注意力输出 $\boldsymbol{\beta}$ 作为补充信息嵌入到 $\boldsymbol{\alpha}$ 中，最后将原始输入 $\boldsymbol{\gamma}$ 作为辅助信息利用残差结构与加权后的特征融合得到增强特征 $F_{re}\in\mathbb{R}^{H\times W\times C}$。鉴于低级特征不同通道之间语义区别很小，且通道注意力在特征通道有一定数量时会发挥更大作用，因此我们在 Layer3 和 Layer4 之间引入该注意力机制，并将模板分支的输出通过该注意力机制来增强模板特征的可辨识度。为了验证本文提出的三元注意力机制的有效性，在 3.3 节将本文注意力机制与其他方法做了对比实验。

2.3　层次特征语义融合

网络对输入图像提取特征的顺序总是由浅至深，通过可视化不同卷积层输出特征，如图 7 所示。可以发现，浅层特征分辨率较高，目标容易辨别，益于对目标定位，深层特征比较抽象，像素块辨识度高，含有强辨别性的语义信息，益于对正负样本分类。

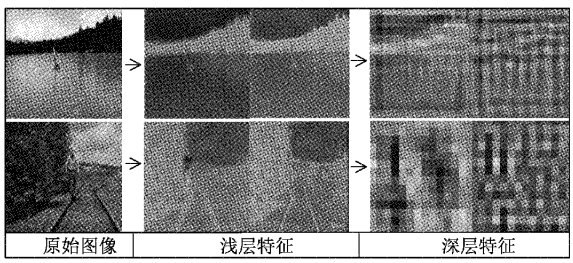
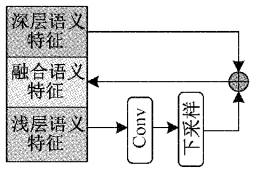


图 7　不同层次特征可视化

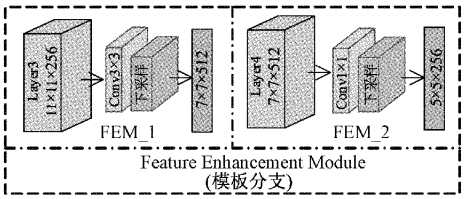
基于不同层次特征有各自优势点这一特性，本文提出一种特征增强模块(feature enhancement module，FEM)，如图 8(a)所示。该模块可以将浅层特征中轮廓、纹理等空间信息与深层特征中的强辨别性信息有效结合，进一步增强特征对目标的表述能力，优化跟踪器在面对复杂跟踪环境时的表现力，所提出的特征增强模块可以表示为：

$$F_{new}=\triangledown(Conv(F_l))\oplus F_h \tag{8}$$

其中，F_h、F_l 分别表示深层特征和浅层特征，$\triangledown(\cdot)$ 表示下采样操作，$Conv$ 为卷积操作，F_{new} 是增强后的特征。具体在网络中实现如图 8(b)模板分支示例，其中，FEM_1 模块利用 3×3 卷积和下采样来调整特征通道维数和分辨率大小，FEM_2 模块借鉴工作[18]利用 1×1 卷积和下采样来实现信息的前向递进融合。



(a) 特征增强模块



(b) 特征增强模块在网络中实现

图 8　层次语义融合

2.4　网络训练

VGG16 网络具有很强的适应性和参数可移植性，因此，本文使用 ImageNet 数据集上预训练的 VGG16 模型(含批量归一化层)参数和 Xavier[19]方法初始化本文的主干网络，并用 GOT-10k[20]训练集进一步训练网络，让网络学习模板图像和搜索图像之间的相似性概念，使其适应跟踪任务。GOT-10k 的训练集包括 560 多种目标类别和 87 种运动模式，含有约 10 000 个视频片段和超过 150 万张有手动标注框的图像。在同一视频片段中随机抽取两张图片分别裁剪成大小为 127×127 和 255×255 来组成模板图像和搜索图像对。整体网络用式(9)损失函数进行训练，将相似得分图的损失定义为单个样本损失取平均值。

$$L(y,v)=\frac{1}{|D|}\sum_{u\in D}\log(1+\exp(-y[u]v[u])) \tag{9}$$

其中，$D\in\mathbb{R}$ 表示相似得分图，u 是 D 上的具体像素位置，$v[u]$ 表示该点的相似度得分，$y[u]$ 表示该点的真实标签，其定义如式(10)所示。

$$y[u]=\begin{cases}+1, & k\|u-c\|\leqslant r\\ -1, & \text{其他}\end{cases} \tag{10}$$

式中：k 是网络总步长，c 表示目标中心，当得分图上的点在以目标中心为圆心，半径为 r 的圆内，则定义为正样本，否则为负样本。

用随机梯度下降(SGD)对训练参数 θ 进行优化更新，如式(11)所示。权重衰减和动量分别设置为 5×10^{-4} 和 0.9，batch size 设置为 8，总计训练 50 个 epoch，用指数衰减的方式将学习率从 1×10^{-2} 下降到 1×10^{-5}。

$$\underset{\theta}{\operatorname{argmin}}\ \underset{(z,x,y)}{E}(L(y,f(z,x;\theta))) \tag{11}$$

3　实验分析

为了验证所提出方法的有效性和鲁棒性，分别在 OTB50，OTB100 和 GOT-10k 等公共基准数据集上对本

文算法进行性能评估,并与一些主流算法进行性能对比。实验平台环境,操作系统为 Windows10,编程语言为 Python3,使用 Pytorch1.8.0 框架,CPU 处理器为 Intel(R) Core(TM) i7-7700,内存为 16 GB,GPU 为 NVIDIA GeForce RTX2060,显存为 12 GB。

3.1　评估指标

OTB[21]和 UAV123[22]数据集的评估指标是跟踪精度和跟踪成功率,其中跟踪精度的指标是中心位置误差,即预测框中心 (x_1,y_1) 与标注框中心 (x_2,y_2) 之间的平均欧氏距离 b,其定义如下:

$$b=\sqrt{(x_1-x_2)^2+(y_1-y_2)^2} \tag{12}$$

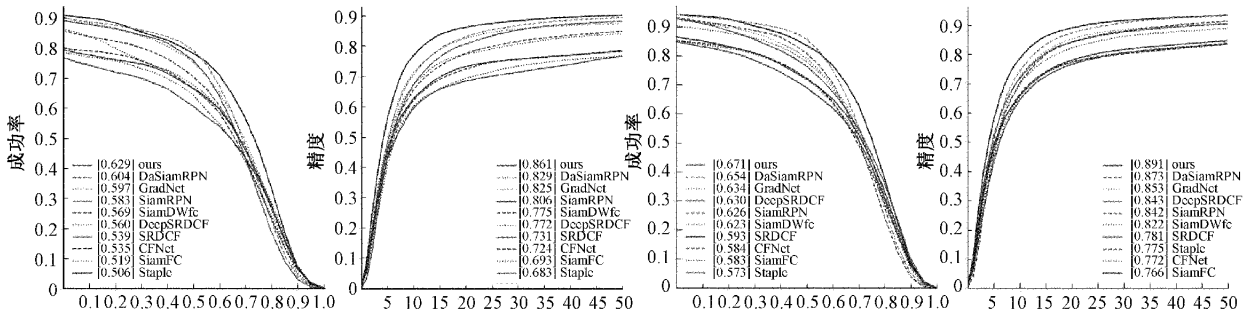
b 越小则精确度越高,取距离在 20 个像素阈值内的视频帧百分比作为最终的跟踪精度。跟踪成功率是通过计算预测框 r_t 和标注框 r_a 的重叠率(intersection-over-union,IoU)在指定阈值下的曲线下面积(area under curve,AUC)得到的,当重合率大于预设阈值时,认为当前帧跟踪成功。IoU 定义如下:

$$IoU=\frac{|r_t\cap r_a|}{|r_t\cup r_a|} \tag{13}$$

GOT-10k 测试集是近几年发布的大规模和多样性的目标跟踪评估数据集,它与训练集没有任何交叉。其评估指标为平均重叠(average overlap,AO)和成功率(success rate,SR),AO 表示所有预测框和真实标注框的平均重叠率,$SR_{0.50}$ 和 $SR_{0.75}$ 分别为阈值在 0.50 和 0.75 下的成功率。

3.2　测试结果

OTB50 和 OTB100 分别包含 50 和 100 个视频序列,这些视频序列涉及到 11 个影响跟踪器性能的属性:光照变化(illumination variation,IV)、尺度变化(scale variation,SV)、形变(deformation,DEF)、遮挡(occlusion,OCC)、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)、超出视野(out-of-view,OV)、背景干扰(background clutters,BC)和低分辨率(low resolution,LR),是跟踪任务种常用的标准评估数据集。在 OTB 数据集上采用一次通过评估(One Pass Evaluation,OPE)将本文算法与 DaSiamRPN[7]、GradNet[23]、SiamRPN[5]、SiamDWfc[8]、SRDCF[24]、CFNet[9]、SiamFC[1]、Staple[25]、Deep-SRDCF[26]等主流算法进行对比实验,实验结果如图 9 所示。从实验结果可以看出,无论是在 OTB50 或 OTB100 数据集上本文算法都取得了十分有竞争性的结果。在(OTB50,OTB100)上本文成功率和跟踪精度较基准算法分别提升(21.1%,15.1%),(24.2%,16.3%)。为了验证本文算法在应对复杂环境时的表现,在 OTB100 数据集中基于 11 种场景对算法性能进行测试,并于其他 9 种算法进行 AUC 对比,结果如表 2 所示,本文算法在复杂场景下也能表现出优异的跟踪性能。



(a) OTB50测试成功率　(b) OTB50测试精度　(c) OTB100测试成功率　(d) OTB100测试精度

图 9　OTB 数据集测试结果

表 2　10 种算法在 OTB100 中 11 种属性的 AUC 对比结果

属性	视频数	Staple	GradNet	SiamRPN	DeepSRDCF	SiamDWfc	SRDCF	SiamFC	CFNet	DaSiamRPN	本文
IV	37	0.579	0.632	0.642	0.608	0.612	0.600	0.563	0.533	0.644	**0.686**
SV	63	0.521	0.614	0.615	0.605	0.613	0.561	0.556	0.533.	0.637	**0.666**
OPR	63	0.525	0.621	0.621	0.599	0.606	0.542	0.555	0.548	0.638	**0.655**
OCC	48	0.533	0.608	0.580	0.591	0.594	0.550	0.542	0.521	0.603	**0.613**
DEF	43	0.540	0.562	0.611	0.555	0.551	0.533	0.505	0.519	0.636	**0.642**
MB	29	0.524	0.631	0.613	0.625	0.641	0.578	0.543	0.529	0.611	**0.695**
OV	14	0.475	0.583	0.542	0.553	0.590	0.460	0.509	0.454	0.537	**0.610**
IPR	51	0.539	0.619	0.622	0.579	0.599	0.534	0.553	0.561	0.644	**0.662**
FM	39	0.528	0.613	0.593	0.611	0.620	0.585	0.562	0.546	0.611	**0.660**
LR	9	0.394	0.669	0.639	0.561	0.596	0.514	0.618	0.614	0.636	**0.712**
BC	31	0.560	0.611	0.591	0.627	0.574	0.583	0.527	0.561	0.642	**0.647**

注:黑体为每行最优,下划线为次优

如图 10 所示，在 GOT-10k 测试集上本文算法较基准算法 SiamFC，AO 从 0.348 增加到 0.431，提升了 23.9%，与去其他主流算法相比也有很大的性能优势。在表 3 中，将本文算法与其他 6 种算法的 $SR_{0.50}$ 和 $SR_{0.75}$ 进行比较，本文算法较基准算法的 $SR_{0.50}$ 和 $SR_{0.75}$ 分别提升 38.2% 和 65.3%。

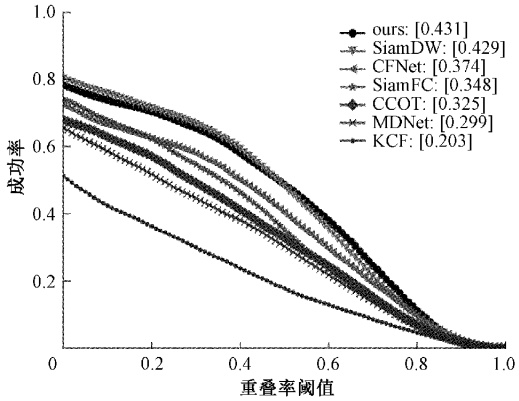


图 10　GOT-10k 测试集结果

表 3　GOT-10k 测试集对比结果

跟踪器	AO	$SR_{0.50}$	$SR_{0.75}$
本文	**0.431**	**0.488**	**0.162**
SiamDW[8]	0.429	0.483	0.147
CFNet[9]	0.374	0.404	0.144
SiamFC[4]	0.348	0.353	0.098
CCOT[27]	0.325	0.328	0.107
MDNet[28]	0.299	0.303	0.099
KCF[29]	0.203	0.177	0.065

注：粗体为每列最优结果，下划线为次优

UAV123 是用于低空无人机目标跟踪的视频数据集，包含 123 个无人机拍摄的高分辨率视频序列。实验结果如表 4，在 UAV123 数据集上，本文算法较基准算法成功率和精度分别提升(11.6%，5.98%)。

表 4　UAV123 数据集对比结果

跟踪器	成功率	精度
本文	**0.556**	**0.762**
SiamRPN[5]	0.540	0.735
ECO-HC[30]	0.506	0.725
SiamFC[4]	0.498	0.719
SRDCF[24]	0.464	0.676
DSST[31]	0.356	0.586

注：粗体为每列最优结果，下划线为次优

3.3　消融实验

为了验证本文所提出的三元注意力机制和层次语义融合两个关键优化特征策略的有效性，以 SiamFC 算法为基准在 OTB100 数据集上设计了消融实验，实验结果如表 5 所示。在改进 VGG16 网络基础上，仅采用层次语义融合策略或三元注意力机制 AUC 分别提升(10.9%，9.4%)，二者结合的情况下 AUC 提升 15.1%。

表 5　本文算法在 OTB100 数据集上的消融实验

算法	主干网络	层次语义融合	三元注意力机制	AUC
SiamFC	AlexNet	×	×	0.583
本文	Modified VGG16	✓	×	0.647
		×	✓	0.638
		✓	✓	0.671

为了进一步验证本文提出的三元注意力机制(Triple-Attn)的有效性，在同等参数配置下，将本文提出的注意力机制与跟踪任务中常用的 SENet，ECA-Net 通道注意力机在 OTB50 数据集上进行性能对比实验，结果如图 11 所示。其中，Base 是结合了层次语义融合模块的 VGG16，在此基础上，提出的注意力机制要比 SENet 等通道注意力表现出更优越的性能。同时，Triple-Attn 是轻量级的，仅有 300 多个参数量，保证跟踪器可以在 40 fps 的速度下实时运行。

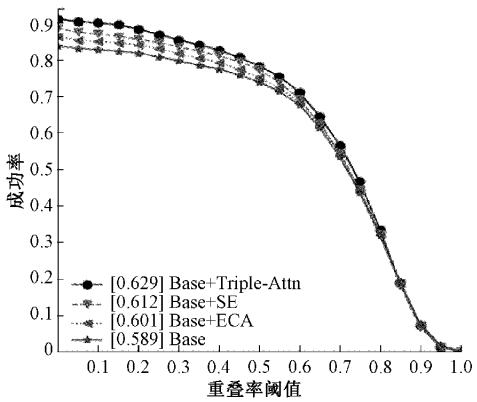
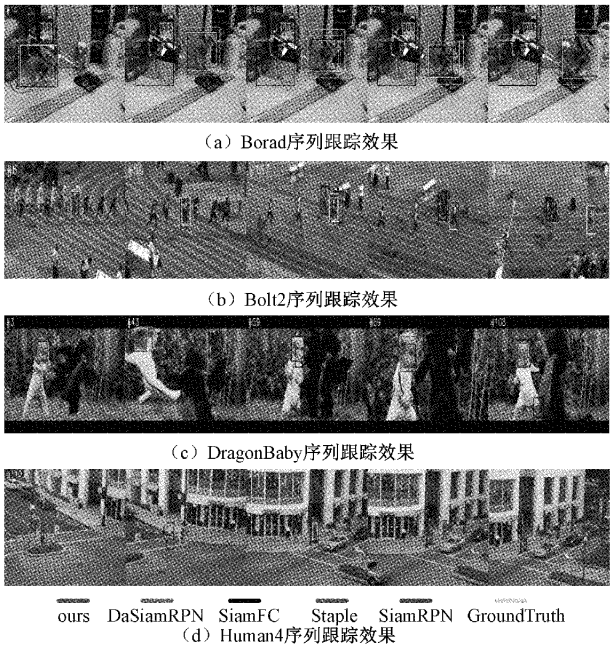


图 11　注意力机制在 OTB50 数据集上的对比结果

3.4　定性分析

本小节主要是通过可视化跟踪效果来直观地对比本文算法与其他算法的性能。选取了本文实验中的 DaSiamRPN，SiamRPN，Staple 和基准算法 SiamFC 在 OTB 数据集中 4 个视频序列(Board，Bolt2，DragonBaby，Human4，)进行跟踪效果对比。这几个视频序列包含相似物体干扰，背景信息复杂，形变，遮挡，光照变化等属性，效果如图 12 所示。图中，每一种颜色框代表一种算法，其中 GroundTruth(黄色)代表真实标注框。a 中，由于背景十分复杂，在第 61 帧的时候 SiamFC 出现严重跟踪漂移，在第 463 帧，目标发生旋转模糊，DaSiamRPN 和 SiamRPN 也发生了跟踪漂移，本文算法依旧对目标进行准确跟踪。b 所

示是目标快速运动且包含相似物干扰因素的场景,在第 187 帧,SiamRPN 丢失目标,在第 215 帧,SiamFC 随之也丢失目标,本文算法始终对目标进行精确跟踪。c 中,在第 43 帧,当目标发生旋转时,Staple,SiamRPN,SiamFC 都出现了跟踪丢失,在第 89 帧和 108 帧,虽然其他算法找回目标,但 SiamFC 彻底丢失目标位置信息。d 所示是对行人小目标进行跟踪,且伴有相似物、光照变化等干扰因素。在第 200 帧时,由于目标周围相似行人的干扰,SiamFC 出现跟踪漂移,虽然在 344 帧时找回目标,但在第 377 帧和 481 帧时,因为出现遮挡,DaSiamRPN,SiamRPN,SiamFC 丢失目标。在以上示例的 4 个包含复杂场景的视频序列中,我们提出的模型预测出的边界框是与真实标注框最为接近的。



(a) Borad序列跟踪效果
(b) Bolt2序列跟踪效果
(c) DragonBaby序列跟踪效果
(d) Human4序列跟踪效果

图 12　5 种算法的跟踪可视化对比

从上述分析可知,本文算法在面对目标遮挡、快速运动、相似物干扰等复杂跟踪环境时,都能始终锁定目标位置。

4　结　　论

针对基于全卷积孪生网络跟踪算法在面对复杂跟踪环境时鲁棒性差的情况,考虑到特征质量在基于孪生网络跟踪模型中的重要性,本文提出了基于强化特征学习和表达策略的孪生网络跟踪算法。通过在 OTB50,OTB100,GOT-10k 等公开基准数据集上实验验证,本文提出的算法具备优秀的跟踪性能,在面对复杂环境时也能保持鲁棒性,同时满足了实时性要求。

参考文献

[1] 严飞,夏金锋,马可,等. 基于 Mean Shift 算法的目标跟踪系统设计[J]. 电子测量技术,2020,43(23):6-11.

[2] 杨梅,贾旭,殷浩东,等. 基于联合注意力孪生网络目标跟踪算法[J]. 仪器仪表学报,2021,42(1):127-136.

[3] TAO R GAVVES E, SMEULDERS A W M. Siamese instance search for tracking[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016: 1420-1429.

[4] BERTINETTO L, VALMADER J, HENRIQUES J F, et al. Fully convolutional siamese networks for object tracking [C]. European Conference on Computer Vision. Springer, Cham, 2016: 850-865.

[5] LI B, YAN J, WU W, et al. High performance visual tracking with siamese region proposal network [C]. Proceedings of the IEEE Conference Recognition, 2018:8971-8980.

[6] REN S, HE K, GIRSHICK R, et al. Faster RCNN towards real-time object detection with region proposal networks [J]. Advances in Neural Information Processing Systems, 2015:1137-1149.

[7] ZHU Z, WANG Q, LI B, et al. Distractor-aware siamese networks for visual object tracking [C]. Proceedings of the European Conference on Computer Vision(ECCV),2018:101-117.

[8] ZHANG Z, PENG H. Deeper and wider siamese networks for real-time visual tracking[C]. Proceeding of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019:4591-4600.

[9] 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, 2017: 2805-2813.

[10] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition [J]. ArXiv Preprint, 2014, ArXiv: 1409-1556.

[11] 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, 2019: 4282-4291.

[12] YU F, KOLTUN V. Multi-scale context aggregation by dilated convolutions[J]. ArXiv Preprint, 2015, ArXiv: 1511.07122.

[13] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7132-7141.

[14] WOO S, PARK J, LEE J Y, et al. Cbam: Convolutional block attention module[C]. Proceedings

of the European Conference in Computer Vision (ECCV), 2018:3-19.

[15] WANG Q, WU B, ZHU P, 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, 2020:11531-11539.

[16] ZEILER M D, FERGUS R. Visualizing and understanding convolutional networks[C]. Proceeding of the European Conference in Computer Vision,2014:811-833.

[17] ZAGORUYKO S, KOMODAKIS N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer[J]. Computer Vision and Pattern Recognition, 2016, DOI:10.48550/arXiv.1612.03928.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778.

[19] HLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks [C]. Proceedings of the Thirteenth International Conference Artificial Intelligence and Statistics, JMLR Workshop and Conference Vision and Pattern Recognition, 2013: 2411-2418.

[20] HUANG L, ZHAO X, HUANG K. GOT-10k: A large high diversity benchmark for generic object tracking in the wild[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019(5):1562-1577.

[21] WU Y, LIM J, YANG M H. Online object tracking: A benchmark [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013: 2411-2418.

[22] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for uav tracking [C]. European Conference on Computer Vision, Berlin German: Springer,2016: 455-461.

[23] LI P, CHEN B, OUYANG W, et al. GradNet: Gradient-guided network for visual object tracking[C]. 2019 IEEE/CVF International Conference on Computer Vision(ICCV), 2019:6161-6170.

[24] DANELLJAN M, HAGER G, KHAN F S, et al. Learning spatially regularized correlation filters for visual tracking [C]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 4310-4318.

[25] BERTINETTO L, VALMADRE J, GOLODETZ S, et al. Staple: Complementary learners for real-time tracking [C]. Computer Vision & Pattern Recognition, IEEE, 2016:1401-1409.

[26] DANELLJAN M, HAGER G, KHAN F S, et al. Convolutional features for correlation filter based visual tracking [C]. Proceedings of the IEEE International Conference on Computer Vision Workshop,2015: 58-66.

[27] DANELLJAN M, ROBINSON A, KHAN F S, et al. Beyond correlation filters: Learning continuous convolution operators for visual tracking[C]. Springer International Publishing,2016:472-488.

[28] NAM H, HAN B. Learning multi-domain convolutional neural networks for visual tracking[J]. IEEE, 2016: 4293-4302.

[29] 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, 2014,37(3):583-596.

[30] DANELLJAN M, BHAT G, SHAHBAZ K F, et al. Eco: Efficient convolution operators for tracking[C]. IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ: IEEE,2017:6638-6646.

[31] DANELLJAN M, HAGER G, KHAN F, et al. Accurate scale estimation for robust visual tracking[C]. British Machine Vision Conference, 2014:1-5.

作者简介

符强,副教授,硕导,主要研究方向为通信与信号处理、图像处理、卫星导航与定位。

E-mail:2325950807@qq.com

王阳,硕士研究生,主要研究方向为目标跟踪。

E-mail:yangyang158969@163.com

纪元法(通信作者),教授,博导,主要研究方向为卫星导航与卫星导航接收机。

E-mail:937315383@qq.com

任风华,副教授,硕导,主要研究方向为图像处理。

E-mail: 2919625683@qq.com