吉林大学学报(工学版)
Journal of Jilin University(Engineering and Technology Edition)
ISSN 1671-5497,CN 22-1341/T

《吉林大学学报(工学版)》网络首发论文

题目：基于多重注意力机制的无锚框目标跟踪算法
作者：刘晶红，邓安平，陈琪琪，彭佳琦，左羽佳
DOI：10.13229/j.cnki.jdxbgxb20220166
收稿日期：2022-02-21
网络首发日期：2022-05-06
引用格式：刘晶红，邓安平，陈琪琪，彭佳琦，左羽佳．基于多重注意力机制的无锚框目标跟踪算法[J/OL]．吉林大学学报(工学版).
https://doi.org/10.13229/j.cnki.jdxbgxb20220166

基于多重注意力机制的无锚框目标跟踪算法

刘晶红[1]，邓安平[1,2]，陈琪琪[1,2]，彭佳琦[3]，左羽佳[1]

(1.中国科学院 长春光学精密机械与物理研究所，长春 130033；2.中国科学院大学，北京 100039；3.中国人民解放军陆军装备部驻沈阳地区军代局驻长春地区第一军代室，长春 130022）

摘要：针对现有孪生神经网络跟踪算法两个分支相互独立缺少信息交互，在受到目标遮挡、相似目标干扰等挑战下无法精确鲁棒跟踪的现状，提出了一种基于多重注意力机制的无锚框目标跟踪算法。本文使用多重注意力机制编码目标模板特征与搜索区域特征，通过自注意力机制提升特征显著性后利用互注意力机制聚合目标模板和搜索区域之间的特征信息，强化了算法对目标与背景的鉴别能力。同时引入无锚框机制，以逐像素的方式完成端到端的视觉目标跟踪任务，避免锚框机制带来的人为干预的弊端。实验结果表明，在 OTB50、OTB100、GOT-10K 公开数据集上，本文提出的基于多重注意力机制的无锚框目标跟踪算法针对目标遮挡以及相似目标干扰等挑战具有较强的鲁棒性，有效提升了跟踪算法的准确率和成功率。

关键词：计算机视觉；目标跟踪；注意力机制；无锚框机制

中图分类号：TP391.4　　**文献标志码：**A

DOI：10.13229/j.cnki.jdxbgxb20220166

Anchor-free target tracking algorithm based on multiple attention mechanism

LIU Jing-hong[1]，　DENG An-ping[1,2]，CHEN Qi-qi[1,2]，PENG Jia-qi[3]，ZUO Yu-jia[1]

（1. Changchun Institute of Optics,Fine Mechanics and Physics,Chinese Academy of Sciences, Changchun 130033,China；2. University of Chinese of Sciences, Beijing 100039, China；3.The First Military Representative Office of the Military Representative Bureau of the Army Equipment Department of the Chinese People's Liberation Army in Shenyang and in Changchun，Changchun 130022,China）

Abstract: Siamese network based trackers have two branches which are independent of each other and lack of information interaction. So it cannot accurately and robust tracking under the challenges of target occlusion and similar object.To solve this problem, an anchor-free target tracking algorithm based on multiple attention mechanism is proposed.In this paper, multiple attention mechanism is used to encode the target template and search area features. After improving the feature significance through selfattention mechanism, mutual attention mechanism is used to aggregate the feature interaction between target template and search area, which strengthens this algorithm’s discrimination ability between target and background. At the same time, the anchorfree mechanism is used to complete the end-to-end visual target tracking task pixel by pixel, avoiding the disadvantages of human intervention caused by the anchor frame mechanism. Extensive experiments are conducted on many challenging benchmarks like OTB-50, OTB-100 and GOT-10K. These results show the anchor-free target tracking algorithm based on multiple attention mechanism proposed in this paper has strong robustness against the challenges of target occlusion and similar object, and effectively improves the precision rate and success rate of the tracking algorithm.

Key words: Computer vision; object tracking; attention mechanism; anchor-free

收稿日期：2022-02-21.
基金项目：国家自然科学基金面上项目（62175233）
作者简介：刘晶红（1967-），女，研究员，硕士. 研究方向：光电成像. E-mail:liu1577@126.com

0 引言

目标跟踪作为一项重要的计算机视觉任务，广泛应用于如智能视频监控、自动驾驶等领域。目标跟踪技术是指在视频首帧中将感兴趣的目标确定为目标模板，分析目标与背景信息，算法在接下来的每一帧中对目标进行状态估计和定位的过程。

目标跟踪在不受约束的环境条件下存在相似目标干扰、严重的遮挡以及目标姿态外观改变等众多挑战。现有主流的孪生神经网络算法两个分支相互独立，无法在计算过程中完成信息交互导致丢失大量可用信息，难以有效应对上述挑战，降低了跟踪任务的准确性以及鲁棒性。针对以上问题，本文设计了一种基于多重注意力机制的无锚框目标跟踪算法，在 SiamCAR[1]的基础上，引入多重注意力机制有效强化并聚合两个分支的特征信息，利用自注意力机制提升各分支的特征显著性；利用互注意力机制强化算法对目标和背景的鉴别能力。

本文贡献如下：

1）提出了一个端到端的无锚框目标跟踪框架，针对相似目标干扰与遮挡等挑战，设计多重注意力机制提升算法跟踪精度与鲁棒性；

2）在特征提取环节，本文通过自注意力机制提升特征显著性，使得特征具有全局上下文信息；进而利用互注意力机制沟通目标模板与搜索区域特征信息使目标特征更具有判别性和表达能力；

3）在 OTB100、GOT-10K 等公开数据集上的优秀性能证明了本文算法的有效性。同时通过运行效率对比实验，证明本文算法在提升跟踪性能的同时并未对运行效率产生影响。

1 相关工作

现有的跟踪算法可划分为生成式模型和判别式模型。生成式模型如光流法[2]，Meanshift 法[3]在出现光照变化、尺度变化、相似物体干扰情况下鲁棒性较差。判别式模型可分为基于相关滤波算法以及基于孪生神经网络算法。基于相关滤波算法中，KCF[4]算法将目标跟踪任务从图像匹配问题转化成区分目标与背景的分类问题，利用循环矩阵扩大样本容量的同时规避矩阵逆运算，增强算法的准确性。

孪生神经网络强大的特征提取能力使得目标跟踪任务精度取得巨大进步，基于孪生神经网络的相关算法引起了目标跟踪领域的广泛关注。孪生神经网络由两个相互独立且权重共享的特征提取网络组成。将目标模板和搜索区域作为两个分支的输入，分别提取相应的特征后通过匹配响应图找到与目标模板相似度最高的搜索区域完成跟踪任务。SINT[5]作为开创性的工作，利用孪生神经网络直接学习匹配函数，通过相似学习解决跟踪问题。SiamFC[6]引入平移等变性，去除图片大小限制，使用 5 个预先设定尺度的锚框回归目标框，然而锚框内含有大量的非必要背景信息导致跟踪精度较低。SiamRPN[7]系列使用区域建议网络将跟踪任务从匹配问题转化为前景与背景的分类问题，通过对同一位置 k 个不同大小比例的锚框进行回归，一定程度上提高了任务精度与鲁棒性。后续研究从特征提取，正负样本比例、特征增强等各方面对 SiamRPN 进行改良[8-10]，但锚框机制为了解决不同尺度与长宽比的任务引入许多超参数，即使在大量调参的同时仍然难以处理物体形变与姿态变化较大的问题，SiamCAR 采用无锚框方法中的 FCOS 思想回归目标框，加强目标特征，抑制背景信息，有效提升了算法跟踪性能。但其孪生神经网络两个分支独立，丢失大量可用背景信息。

以 scSE[11]、Non-local Net[12]以及 CBAM[13]为代表的注意力机制具有增强网络特征显著性的能力，从而推进了深度学习的发展。SA-Siam 将通道注意力机制引入语义网络中，增强特征语义信息[14]。RASNet 将残差注意力机制、通用注意力机制以及通道注意力机制三者结合，增强特征的显著性[15]。但只对单个分支使用，精度与鲁棒性仍待提高。才华等[16]提出动态模板更新的策略，在应对目标姿态外形变化时鲁棒性较好，但跟踪过程中收集的不可靠正样本可能导致模型退化。现有算法仍将孪生神经网络中两个分支进行独立计算，无法在计算过程中完成信息交互导致丢失大量可用信息，从而降低了跟踪任务的准确性以及鲁棒性。

2 本文算法

本文算法的整体框架如图 1 所示。采用共享特征提取网络权重的两个分支提取目标模板与搜索区域的深浅层特征，进而分别使用多重注意力机制提升特征显著性并沟通两个分支的特征信息，将目标模板与搜索区域特征进行深度可分离互相关运算，并将响应图深浅层特征融合，最终使用无锚框机制的分类回归网络得到目标边界框。

2.1 特征提取骨干网络

本文使用改进的 ResNet-50 网络[17]作为特征提取骨干网络。为了使特征提取网络更适用于目标跟踪任务，本文尽可能减少卷积中的填充防止信息污染，只在 ResBlock 中间层增加[1,1]的填充；并将特征提取网络的空间步长设置为 1 以保留更多的特征信息。本文采用通道级联的方式将 ResBlock 的三个输出作为特征提取网络的输出。

$$\theta(x) = Concat(f_3, f_4, f_5) \tag{1}$$

其中 (f_3, f_4, f_5) 对应 ResBlock 最后三层的输出特征，

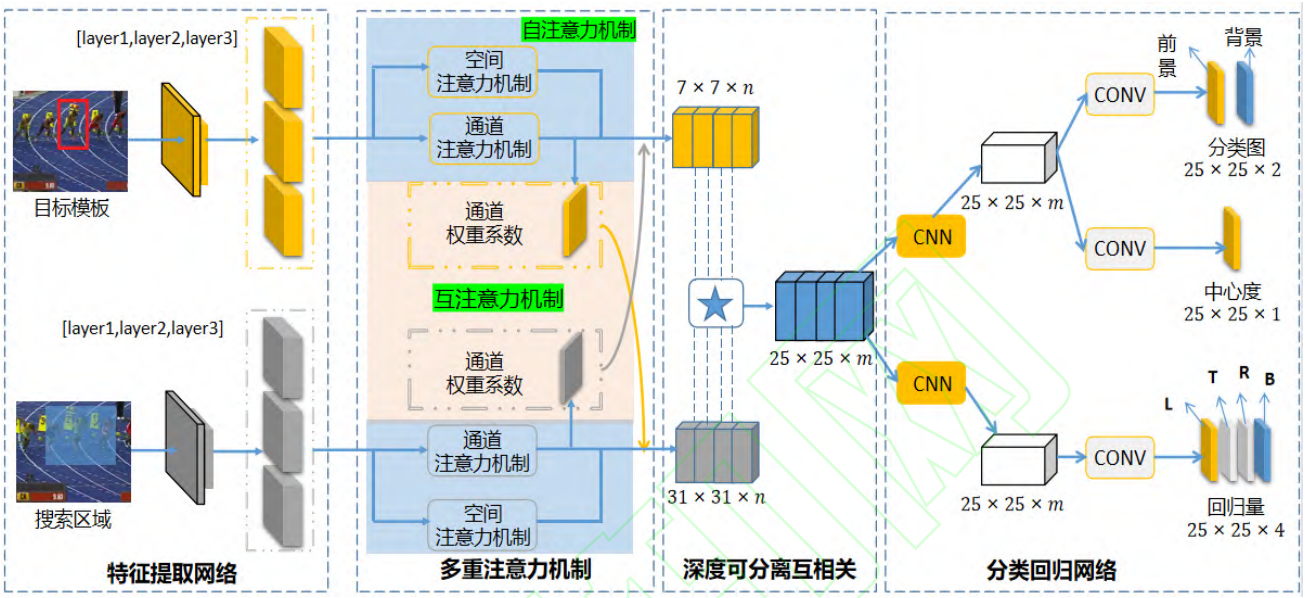


图 1 本文算法框架图

Fig.1 The algorithm frame diagram of this paper

通道数为[512,1024,2048]。为了减少参数数量与计算复杂度利用 1*1 卷积将各层均调整为 256 个通道数。通过通道级联方法将各层连接，最终输出通道数为 3*256。

2.2 多重注意力机制

本文设计的多重注意力机制主要包含自注意力机制以及互注意力机制，利用自注意力机制提升特征显著性进而采用互注意力机制沟通两个分支的特征信息。

2.2.1 自注意力机制

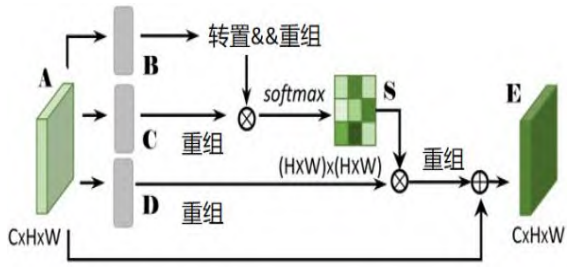


图 2 空间注意力机制图

Fig.2 Spatial attention mechanism map

自注意力机制包含空间注意力机制以及通道注意力机制。对于空间位置，由于网络感受野的限制使得网络特征信息只能捕捉到局部邻域信息，无法利用全局上下文信息，导致网络在应对目标、出视野以及尺度变化时鲁棒性较差。本文引入空间注意力机制使网络学习如何利用全局上下文信息。空间注意力机制如图 2 所示。

将图像的深度特征作为该网络的输入特征 A_{C*H*W}，其中 C 表示特征通道数，H 与 W 分别为特征图的高与宽。利用不同的 1*1 卷积分别得到 B、C、D 并将其大小重组为 $C \times N$ 其中 N 为特征图高与宽相乘，代表特征图像素总个数。使用 Softmax 算子得到空间注意力系数 S，公式如下。

$$S \in R^{N \times N} : S_{ji} = \frac{\exp(B_i \bullet C_j)}{\sum_{i=1}^{N} \exp(B_i \bullet C_j)} \tag{2}$$

式中 S_{ij} 代表第 i 个位置对第 j 个位置的影响程度，即第 i 个位置与第 j 个位置的相似性，数值越大相似程度越高。将空间权重系数重组后与 D 相乘并赋予权重系数 α，再重组为输入特征相同的大小，

最终与原始特征 A_{C*H*W} 相加得到输出 E，公式如下。

$$E \in R^{C\times H\times W} : E_j = \alpha\sum_{i=1}^{N}(S_{ji}\bullet D_i) + A_j \qquad (3)$$

其中 α 代表权重系数，其初始为 0.001，通过本文第 3 节的训练策略，使网络学习得到合适的权重。输出 E 是所有位置相互影响结果与原始位置信息的加权和，因此输出 E 具有全局上下文信息，在保留原始位置信息的同时使用空间注意力权重系数选择性聚合上下文信息。

在神经网络中，每一个通道的响应通常代表某种特定类别的特征，而在跟踪任务中，目标模板在第一帧给出，代表某一种特定类别且保持不变。大多数通道响应对应的特征类别与目标模板不同，平等地对待每一个通道会限制网络的表达能力。因此本文引入通道注意力机制，通过计算通道权重系数自适应地对通道进行加权，增强相关通道的特征显著性并减少不相关通道的影响，如图 3 所示。

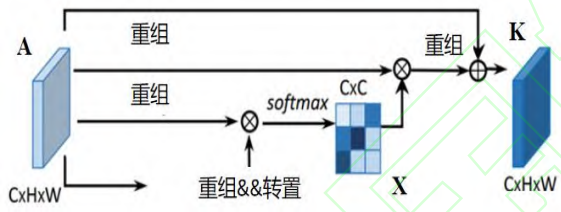


图 3 通道注意力机制图

Fig.3 Channel attention mechanism map

将图像的深度特征作为该网络的输入特征 A_{C*H*W}，其中 C 为特征通道数，H 与 W 分别为特征图的高和宽。通过重组、转置以及 Softmax 算子得到通道权重系数，公式如下。

$$X \in R^{C\times C} : X_{ji} = \frac{\exp(A_i\bullet A_j)}{\sum_{i=1}^{C}\exp(A_i\bullet A_j)} \qquad (4)$$

式中 X_{ji} 代表第 i 个通道对第 j 个通道的影响程度，数值越大影响程度越大。将通道权重系数重组为输入特征相同大小，乘上权重系数 β，并与输入特征 A_{C*H*W} 相加，得到输出特征 K，公式如下。

$$K \in R^{C\times H\times W} : K_j = \beta\sum_{i=1}^{N}(X_{ji}\bullet A_i) + A_j \qquad (5)$$

式中权重系数 β 初始为 0.001，通过本文第 3 节的训练策略，使网络学习得到合适的权重。通道响应 K 是所有通道响应相互影响程度与原始通道响应的加权和。通过元素求和以及 3*3 卷积方法对空间注意力机制输出 E 与通道注意力机制输出 K 进行特征融合。最终得到的输出具有丰富的全局上下文信息以及较强的特征显著性。

2.2.2 互注意力机制

在孪生神经网络当中，目标模板分支与搜索区域分支相互独立，无法有效利用大部分背景信息，而这些背景信息在区分目标与相邻背景时十分重要。在跟踪任务中，存在目标遮挡、相似目标干扰以及目标外观剧烈变化的情况，容易造成跟踪漂移。因此，本文引入互注意力机制，将目标模板与搜索区域两个分支信息实时进行交互，保留大量可用背景信息以有效区分目标与背景。如图 4 所示。

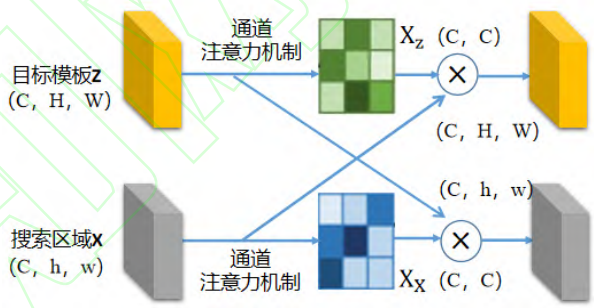


图 4 互注意力机制图

Fig.4 Cross attention mechanism map

采用上文所述通道注意力机制得到目标模板与搜索区域各自对应的通道权重系数。对于搜索区域分支，与模板分支的通道权重系数相乘，使得搜索分支学习得到目标信息，更有效区分目标与背景信息，产生更强有力的特征表达。在应对目标遮挡以及相似目标干扰情况时，算法能更准确识别目标。对于模板分支，与搜索分支的通道权重系数相乘，使得模板分支能学习得到搜索区域中的上下文信息，达到在线隐式更新目标模板的作用。

最后采用元素求和方式完成多重注意力机制，将自注意力机制以及互注意力机制结果相加，得到最终的特征信息。该特征具有整体空间上下文信息，对感兴趣目标与背景具有更好的判别能力，同时完成目标模板的隐式更新。

2.3 深度可分离互相关与深浅层特征融合

互相关运算是指以目标模板为核，与搜索区域进行相关运算得到特征响应图的过程。本文受文献[18]启发，引入深度可分离互相关。深度可分离互相关是将搜索区域作为输入，目标模板作为卷积核，

对二者做逐通道卷积得到目标模板在搜索区域上的互相关响应图。在大幅降低参数量的同时保证跟踪精度不受影响。

在神经网络中，浅层特征能更好的表示如颜色、边缘、尺度和角点等空间信息有利于目标定位以及边界框尺寸和长宽比的确定，深层特征更多关注类别、属性等语义信息更有利前景与背景的区分。有效融合浅层和深层特征有助于提升跟踪精度，因此本文对深度可分离互相关后得到的特征响应图进行特征融合。由上文可知，响应图此时通道数为 3x256，分别代表 ResNet-50 最后三个 ResBolck 输出特征图。本文利用 1*1 卷积将特征响应图降维至 256 个通道，将深浅层特征信息进行融合。利用卷积融合的方法，通过训练使网络自适应更新权重，学习深浅层特征最有效的融合方法。

2.4 分类与回归网络

响应图中每一个位置都可以映射至搜索区域对应位置。基于区域建议网络的算法找到响应图最大值对应搜索区域位置，以此作为多尺度锚框的中心，对锚框位置与尺寸进行回归训练。该方法使用多尺度锚框引入大量超参数，需要先验信息对锚框尺寸进行初步设定。因此本文采用基于无锚框机制的分类回归网络，直接对每个位置的候选框进行分类与回归，避免先验信息的干预以及冗余的超参数调整。

跟踪任务被分解为两个子分支：分类分支以及回归分支。分类分支区分目标与背景，预测该位置的类别。回归分支计算该位置的目标边界框。在网络中，分类分支利用卷积输出大小为 25*25 且通道数为 2 的分类特征图。分类特征图中的两个通道分别表示该位置前景与背景置信度得分。如果分类特征图上位置 (w,h) 所对应搜索区域上位置 (i,j) 落在真实边界框内，则该位置为前景，否则为背景。回归分支输出大小为25*25通道数为4的回归特征图，其中四个通道表示对应搜索区域位置距离真实边界框四条边的距离，记作 (l,t,r,b)，计算公式如下。

$$t_0(i,j)=l=x-x_0 \qquad t_1(i,j)=t=y-y_0 \qquad (6)$$
$$t_2(i,j)=r=x_1-x \qquad t_3(i,j)=b=y_1-y$$

其中 (x,y) 代表在点 (i,j) 对应的搜索区域的位置坐标，(x_0,y_0) 与 (x_1,y_1) 代表真值框左上角和右下角坐标。在实验过程中，发现远离目标中心位置的边界框会产生低质量的预测边界框，影响跟踪精度。因此本文引入一个与分类分支平行的中心度分支去除异常值，生成大小为 25*25 通道数为 1 的向量，表示对应位置的中心度得分 $C(i,j)$。利用公式计算中心度得分，其中当样本点在真值框内时 I =1 否则为 0。

$$C(i,j)=I\times\sqrt{\frac{\min(l,t)}{\max(l,t)}\times\frac{\min(r,b)}{\max(r,b)}} \qquad (7)$$

为了优化分类与回归网络，本文对分类分支采用二元交叉熵损失函数计算 L_{cls}，定义公式如下。

$$L_{cls}=0.5\times BCELoss(\delta_{pos},I)+0.5\times BCELoss(\delta_{neg},I) \quad (8)$$

其中 δ_{pos} 与 δ_{neg} 代表该位置的前景与背景得分，I 代表该位置的标签值，$BCELoss$ 为常见二元交叉熵损失函数。利用 L_{reg} 代表分类分支的损失函数，采用 IOU 损失函数进行计算，公式如下。

$$L_{reg}=\frac{1}{\sum_{(i,j)}I}\sum_{(i,j)}I\times IOU[T(i,j),t(i,j)] \qquad (9)$$

其中 $T(i,j)$ 代表真实坐标对应搜索区域位置距离真实边界框四条边的距离，$t(i,j)$ 代表预测坐标对应搜索区域位置距离真实边界框四条边的距离，IOU 为常见交并比损失函数。本文采用二分类交叉熵损失函数对中心度分支进行优化，公式如下。

$$L_{cen}=\frac{-1}{\sum_{(i,j)}I}\sum_{I=1}C(i,j)\times\log Z(i,j)+(1-C(i,j)\times\log(1-Z(i,j)) \qquad (10)$$

其中 $C(i,j)$ 预测某个位置的中心度得分，$Z(i,j)$ 表示该位置实际中心度得分。本文算法总损失函数如下。

$$L=L_{cls}+\alpha_1 L_{cen}+\alpha_2 L_{reg} \qquad (11)$$

L_{cls}，L_{cen}，L_{reg} 分别代表分类损失函数，中心损失函数以及回归损失函数，利用 α_1 和 α_2 作为权重超参数对网络进行调整，通过实验[1]和经验在训练过程中分别设置为 1 和 3。

3 实验结果与分析

3.1 实验环境

本文算法均在 Windows 系统环境下运行，采用 Python3.7+pytorch1.10 编程框架训练并验证算法性

能，利用 PySOT 库执行算法的训练以及验证过程。本文实验均是在 AMD 3600X 和 GeForce RTX3060ti 显卡上运行，测试过程平均帧率为 35 帧每秒。

3.2 数据集

本文使用 COCO[19],GOT-10K[20],VID 数据集作为训练集训练网络，样本数量为每轮 300000 帧。利用 ResNet-50 在 ImageNet[21]上的训练结果作为主干特征提取网络的预训练模型。本文使用 GOT-10K、OTB50 以及 OTB100[22]数据集作为测试集。OTB50 包含 50 个视频序列、OTB100 数据集有 100 个视频序列、GOT-10K 包含不同目标类别和动作类别的 180 段视频。

3.3 训练过程

本文采用 ResNet-50 在 ImageNet 数据集上的训练结果作为主干网络的预训练模型。在训练过程中采用随机批量梯度下降对模型进行训练，批次大小为 32。本文使用权重衰减改变学习率，前 5 个迭代过程利用热身训练[17]，其中初始学习率为 0.001 随后每个迭代过程增加 0.001，热身结束后采用学习率梯度下降对网络进行训练。共计 20 轮迭代过程，网络总体训练时间为 50 小时。

3.4 OTB50 基准实验结果

OTB50 数据集由 50 个完全标注的视频组成，并提供 11 种视频标注属性，如光照变化、遮挡、尺度变化等，且每一帧图像至少包含两种视频标注属性。本文算法在 OTB50 数据集上进行定量分析以及定性分析。本文算法采用准确率以及成功率作为跟踪性能评价指标。跟踪的准确率是指中心定位误差 CLE 小于 20 像素大小的帧数量占总帧数的百分比。涉及算式如下。

$$CLE = \sqrt{(x_{pr}, y_{pr})^2 + (x_{gt}, y_{gt})^2}$$

$$f = \begin{cases} 1 & CLE \le 20 \\ 0 & CLE > 20 \end{cases} \tag{12}$$

$$precision = \frac{\sum_{i=1}^{N} f}{N}$$

其中 (x_{pr}, y_{pr}) 和 (x_{gt}, y_{gt}) 分别表示算法预测框与真值框的中心位置坐标，中心定位误差 CLE 是指两个中心位置的欧式距离。跟踪算法的成功率是指重叠分数 S 大于 0.5 的帧数量占总帧数的百分比。重叠分数是指计算序列每一帧的预测框 R_{pr} 与真实框 R_{gt} 的面积交并比，取所有帧的交并比之和的平均值。涉及算式如下。

$$S = IOU \frac{|R_{pr} \bigcap R_{gt}|}{|R_{pr} \bigcup R_{gt}|}$$

$$f = \begin{cases} 1 & IOU \le 0.5 \\ 0.5 & IOU > 0.5 \end{cases} \tag{13}$$

$$AUC = \frac{1}{N} \sum_{i=1}^{N} f$$

3.4.1 定量分析实验

本文算法与现有主流算法在 OTB50 数据集上进行对比实验，结果如图 5 所示。

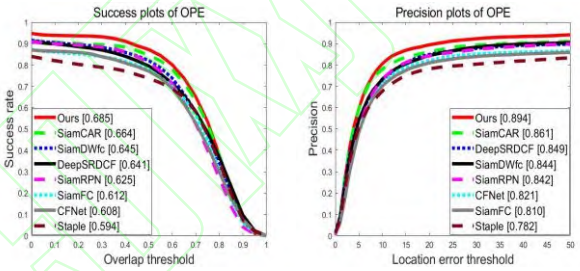
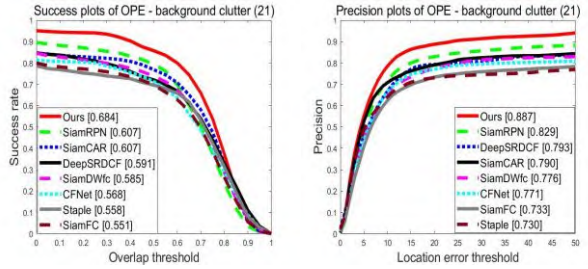


图 5 OTB50 对比图

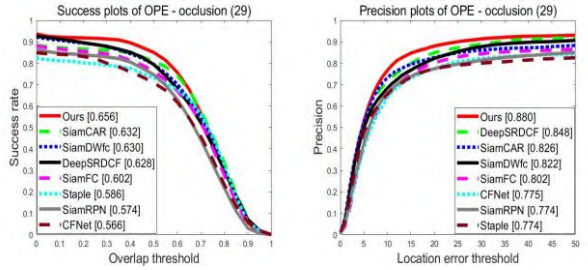
Fig.5 OTB50 Comparison chart

本文算法与对比算法在 OTB50 数据集上进行对比实验，结果如图 5 所示。从图中可以看出，本文算法的精确度和成功率分别达到 68.5%和 89.4%。与对比算法相比，在成功率上存在 2%至 9%提升，在准确率上存在 3%至 10%提升。最重要的是与基准算法 SiamCAR 相比，本文的精确度和成功率分别提升了 2.1%和 3.3%。这表明本文提出的基于多重注意力机制的无锚框目标跟踪算法是有效的。同时，本文算法在 OTB50 数据集上跟踪速度能达到 45 帧每秒，可以实现鲁棒的实时跟踪目标。

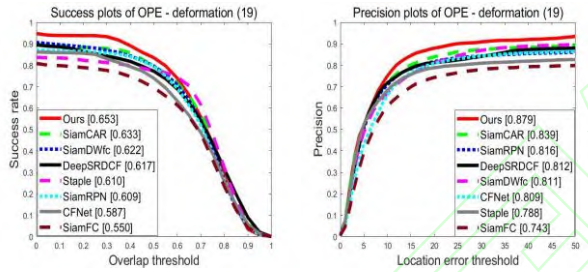
为进一步说明本文算法在应对相似目标干扰（Background clutter，BC）、目标遮挡（Occlusion，OCC）、物体变形（Deformation，DEF）、物体旋转（Out-of-Plane Rotation，OPR）等挑战时的表现，与对比算法在 OTB50 数据集上对这四个挑战属性进行了对比分析。如图 6 所示。



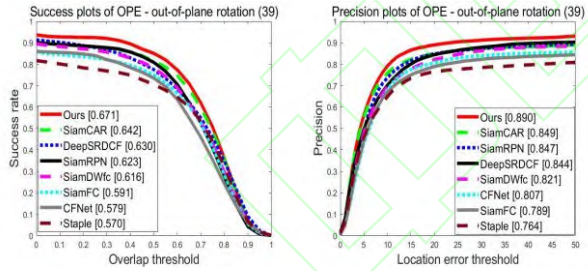
（a）相似目标干扰



（b）目标遮挡



（c）物体变形



（d）物体旋转

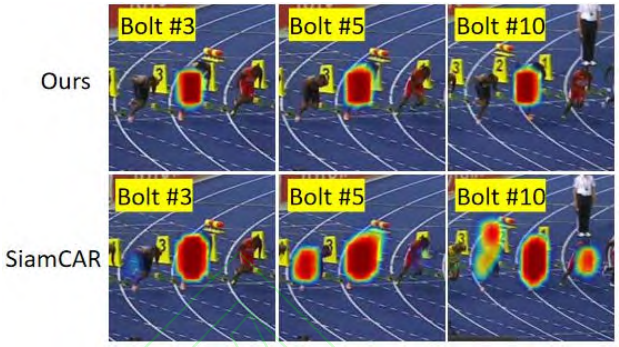
图 6 OTB50 上四个挑战属性的准确率和成功率对比图

Fig.6 Comparison of accuracy and success rate of 4 attributes on OTB50 dataset

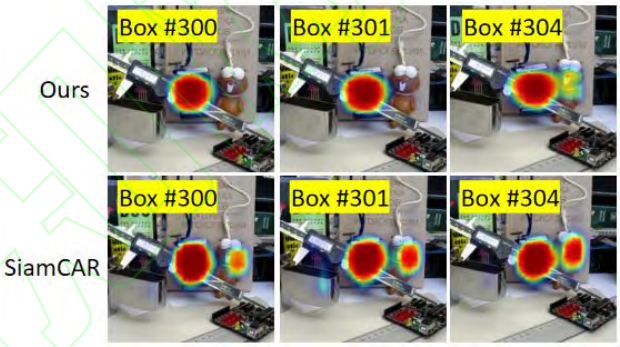
图 6 表明，相较于基准算法，本文算法在相似目标干扰下准确率和成功率提升 7.7%和 9.7%，提升幅度较大。在目标遮挡、物体变形和物体旋转挑战下准确率和成功率均有 4%-6%的提升。这是因为本文算法的特征具有空间全局上下文信息并突出目标特征显著性，因此在应对物体空间信息变化以及外形变化时具有较强的鲁棒性。上述实验进一步表明本文算法的多重注意力机制有效提升算法鲁棒性。

3.4.2 定性分析实验

为了展示本文算法跟踪结果，进一步说明算法在应对目标被遮挡以及相似目标干扰情况下的性能。首先采用热力图形式展现算法感兴趣目标。如图所示。



（a）Bolt 视频序列



（b）Box 视频序列

图 7 算法热力图对比图

Fig.7 Comparison of heat map

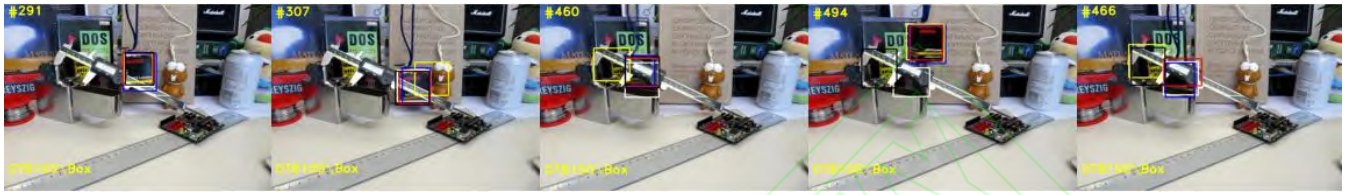
在 Bolt 序列中，该视频序列存在相似目标干扰以及目标快速运动挑战，通过热力图可以发现使用多重注意力机制的本文算法可以明显的区分目标与相似目标而基准算法 SiamCAR 受相似目标干扰影响较大。在 Box 序列中，存在着目标被部分遮挡的现象，通过热力图可以发现本文算法聚焦于目标本身未受遮挡影响，而基准算法 SiamCAR 受遮挡影响较大出现注意力偏移。

为了直观展示本文算法跟踪结果，进一步说明算法在应对目标旋转、遮挡以及相似目标干扰情况下的性能，本文在 OTB100 数据集中选取 5 组具有上述挑战性质的视频序列，与 SiamCAR、SiamRPN++以及 SiamRPN 这 3 个具有代表性的主流跟踪算法进行定性分析实验。在图 6 中、蓝色框代表真值框，红色框代表本文算法，黄色框、黑色框以及白色框分别代表 SiamCAR、SiamRPN++以及 DaSiamRPN 算法结果。

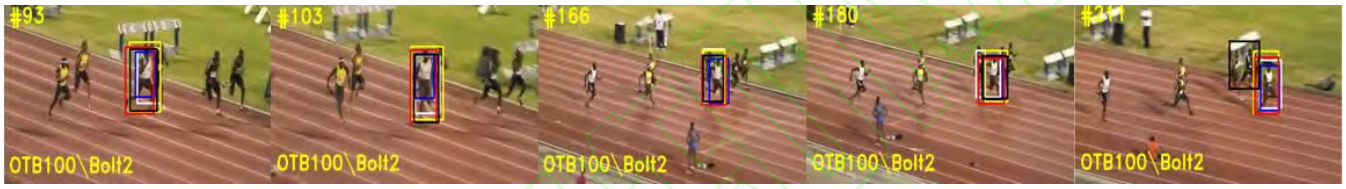
如图 8（a）所示，在 Box 视频序列中，其主要

难点在于存在选定目标被部分遮挡同时目标存在旋转运动的情况。在第 291 帧至 466 帧中，目标在被游标卡尺遮挡情况下不断运动，其中 307 帧时基准算法 SiamCAR 算法发生跟踪漂移，该现象与上文热力图对比结果相一致。在后续帧中，SiamRPN++与 DaSiamRPN 均存在部分跟踪漂移现象。即使在跟踪结果正确时，本文算法仍与目标尺度匹配度最高。由此可以看出，本文算法引入多重注意力机制后，在出现目标被遮挡情况时仍可以鲁棒精准跟踪。

如图 8（b）所示，在 Bolt2 视频序列中，其主要难点在于存在大量相似目标干扰并且物体始终快速运动。在 93 帧与 103 帧中，基准算法 SiamCAR 虽然仍跟踪目标但跟踪框将相似目标也包裹在内，跟踪框尺寸出现较大偏差并且存在跟踪漂移可能。在后续帧中，DaSiamRPN 也存在上述现象，SiamRPN++在 211 帧中受相似目标影响彻底跟丢目标，而本文算法能将目标完整包裹，在表述性上比真值框更加准确。由此可以看出，本文算法在使用



(a) Box 视频序列



(b) Bolt2 视频序列

(c) Skating2 视频序列

(d) Suv 视频序列

(e) Girl2 视频序列

图 8 定性结果对比图

Fig.8 Comparison chart of qualitative results

多重注意力机制能良好的分辨目标与相似背景，在相似目标干扰以及物体快速运动状况下仍能持续稳定跟踪。

如图 8（c）所示，在 Skating1 视频序列中，其主要难点在于存在相似目标干扰以及目标快速旋转。在 107 帧至 195 帧中，目标始终旋转且周围始终伴

随相似目标的干扰，在 175 帧中基准算法 SiamCAR 算法受上述影响出现跟踪漂移现象，后续帧中 SiamRPN++算法与 DaSiamRPN 算法也出现跟踪失败。本文算法始终正确跟踪目标，由此可以说明，本文算法利用多重注意力机制将目标与背景信息建立联系，隐式更新目标模板，有效应对目标旋转以及相似目标干扰情况。

如图 8（d）所示，在 Suv 视频序列中，该视频序列主要难点在于目标被树木大范围长时间遮挡。在 516 帧目标出现被遮挡情况时，基准算法 SiamCAR 已经出现跟踪框不匹配现象，在 530 帧至 550 帧中，基准算法 SiamCAR、SiamRPN++以及 DaSiamRPN 出现跟踪失败，而本文算法虽然仍能跟踪目标但跟踪框尺寸出现偏差，可能是受长时间遮挡影响。在 559 帧中，目标再次出现时，本文算法仍能较正确跟踪。由此可以说明，本文算法可以良好应对目标长时间遮挡情况。

如图 8（e）所示，在 Girl2 视频序列中，该视频序列难点在于目标存在被相似目标所遮挡情况。在 105 帧中目标开始被相似目标遮挡时，基准算法 SiamCAR 出现跟踪漂移现象。在 108 帧至 131 帧中相似目标不断运动并远离跟踪目标，SiamCAR、SiamRPN++以及 DaSiamRPN 算法均出现跟踪失败现象，本文算法在目标再次出现后仍能正确跟踪目标。在最后两帧中，目标与相似目标再次相遇，本文算法仍保持着与真值框的高重叠率，而其余算法被相似目标所影响，跟踪框重叠率较低。

上述定性实验均进一步证实本文算法提取的特征具有空间全局上下文信息并突出目标特征显著性，因此在应对背景干扰以及外形变化时具有较强的鲁棒性。表明本文算法具有一定先进性。

3.5 GOT-10K 基准实验结果

GOT-10K 数据集共包含超过 1 万条视频可分为 560 多个类别。为了检验模型的泛化能力，在训练集与测试集中不存在类别相同情况。将本文算法与基准算法 SiamCAR 以及 SiamRPN++、DaSiamRPN、SiamRPN、ATOM 等当今主流跟踪算法进行对比。采用平均重叠率(Average Overlap)、阈值不同时的准确率（$SR_{0.5}$、$SR_{0.75}$）以及帧率（FPS）作为评价指标。如表 1 所示。

表 1 GOT-10K 实验对比表

Table 1 GOT-10K comparison chart

算法	SiamFC	SiamRPN	SiamRPN++	ATOM	SiamCAR	Ours
AO	0.374	0.463	0.516	0.556	0.569	0.576
$SR_{0.50}$	0.404	0.549	0.620	0.634	0.670	0.672
$SR_{0.75}$	0.144	0.253	0.334	0.402	0.415	0.439
FPS	25.8	74	26	21	18	17

表 1 是各算法在 GOT-10K 上的实验结果，$SR_{0.75}$ 是指预测框与真值框的重合面积比超过 0.75 的帧占比，$SR_{0.5}$ 同理。本文算法具有较高的平均重叠率，在 $SR_{0.5}$ 中相较于基准算法 SiamCAR 存在一定进步（0.7%），在 $SR_{0.75}$ 中相较于基准算法 SiamCAR 具有较大提升（2.4%）。

这是因为本文算法采用多重注意力机制提升特征显著性、聚合目标模板和搜索区域之间的特征信息提升模型抗背景干扰及相似干扰能力，使模型更关注目标本身，最终得到的跟踪框更贴合目标本身，因此 $SR_{0.75}$ 较高。实验进一步说明本文采用的多重注意力机制的有效性。同时，本文在提升精度的同时仍能保持与基准算法相近跟踪速度。

3.6 运行效率对比实验

本节将对多重注意力机制对算法运行效率的影响进行评估。其中 FLOPs 代表计算量，用来衡量算法的复杂度；Params 代表模型参数总量，指模型训练中需要训练的参数总数；FPS 代表算法每秒运行的帧数，以 GOT-10K 作为测试集。结果如表 2 所式。

从表 2 中可以看出，当 127×127×3 与 255×

255×3 分别作为目标模板与搜索区域大小时，本文算法引入多重注意力机制后，相较于基准算法 SiamCAR 计算量上提升 1.3G、参数量上提升 2.0M 为原有参数量的 102%、FPS 降低 1。这说明本文采用的多重注意力机制，不仅能有效提升目标跟踪精度以及有效改善算法应对目标遮挡以及相似目标干扰，同时对算法的运行效率影响较小。

表 2 运行效率对比实验表

Table 2 Comparison chart of efficiency

算法	FLOPs	Params	FPS
SiamCAR	83.2G	91.9M	18
Ours	84.5G	93.9M	17

4 结束语

综上所述，本文提出了一种基于多重注意力机制的无锚框目标跟踪算法。针对 SiamCAR 基准算法在应对目标遮挡、相似目标干扰等复杂挑战下的局限性，利用多重注意力机制捕获丰富的空间上下文信息并在通道域选择性增强相互依赖的通道特征，进而通过互注意力机制聚合目标模板与搜索区域有意义的上下文信息，达到在线隐式更新模板的目的。通过全面的实验：本文算法在 OTB50 数据集上具有较优的准确率和成功率；在 OTB100 数据集上进行热力图分析以及定性实验分析，算法更关注目标本身；在 GOT-10K 数据集上与对比算法相比较平均重叠率等指标均有提升。实验结果表明，本文算法具有较强的跟踪精度与鲁棒性，有效解决了目标遮挡、相似目标干扰引起的跟踪漂移等问题，且对算法运行效率影响较小，具有良好的应用价值。

参考文献：

[1] Guo D, Wang J, Cui Y, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6269-6277.

[2] Baker S, Matthews I. Lucas-kanade 20 years on: A unifying framework[J]. International journal of computer vision, 2004, 56(3): 221-255.

[3] Collins R T. Mean-shift blob tracking through scale space[C]//2003 IEEE Computer Society conference on computer vision and pattern recognition, 2003. Proceedings. IEEE, 2003, 2: II-234.

[4] Henriques J F, Caseiro R, Martins P , et al. High-Speed Tracking with Kernelized Correlation Filters[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2015, 37(3):583-596.

[5] 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.

[6] Bertinetto L, Valmadre J, Henriques J F, et al. Fully-convolutional siamese networks for object tracking[C]//European conference on computer vision. Springer, Cham, 2016: 850-865.

[7] Li B, Yan 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. 2018: 8971-8980.

[8] Li B, Wu W, Wang Q, et al. Evolution of siamese visual tracking with very deep networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, Long Beach, CA, USA. 2019: 16-20.

[9] 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.

[10] 王侃,苏航,曾浩,等.表观增强的深度目标跟踪算法[J/OL].吉林大学学报(工学版):1-10[2022-02-16].https://doi.org/10.13229/j.cnki.jdxbgxb20210367.

Wang K, Su H, Zeng H,et al.Deep target tracking using augmented apparent information.[J/OL].Journal of Jilin University(Engineering Edition):1-10[2022-02-16].https://doi.org/10.13229/j.cnki.jdxbgxb20210367.

[11] Roy A G, Navab N, Wachinger C. Concurrent spatial and channel ‘squeeze & excitation’ in fully convolutional networks[C]//International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2018: 421-429.

[12] Wang X, 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.

[13] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on

computer vision (ECCV). 2018: 3-19.

[14] He A, Luo C, Tian X, et al. A twofold siamese network for real-time object tracking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4834-4843.

[15] Wang Q, Teng Z, Xing J, 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. 2018: 4854-4863.

[16] 才华,王学伟,朱新丽,等.基于动态模板更新 S 的孪生网络目标跟踪算法[J/OL].吉林大学学报(工学版):1-12[2022-02-16].DOI:10.13229/j.cnki.jdxbgxb20200962.

Cai H, Wang X, Zhu X,et al.Siamese network target tracking algorithm based on dynamic template updating[J/OL].Journal of Jilin University(Engineering Edition):1-12[2022-02-16].DOI:10.13229/j.cnki.jdxbgxb20200962.

[17] 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.

[18] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[19] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[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, 43(5): 1562-1577.

[21] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[22] 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.