文章编号:1007-2780(2024)11-1569-12

基于Swin-Transformer改进的目标跟踪算法

刘　时,朱　明*
(中国科学院 长春光学精密机械与物理研究所,吉林 长春 130033)

摘要:基于STARK目标跟踪方法中采用ResNet为骨干网络,其特征提取能力不足,跟踪效果较差。针对此问题,本文基于Swin-Transformer网络,提出了一种改进的目标跟踪算法。首先,对Swin-Transformer内窗口注意力机制进行多尺度改进,设计多尺度窗口模块MW-MSA,旨在提取更为丰富的局部细节信息,与全局上下文信息共同构成多尺度判别性特征。接着,结合Transformer的编码-解码结构作为特征融合网络,采用优化的多层感知机作为更新分数判断网络构成状态感知模块。最后,针对目标消失、重现挑战,提出了一种多跟踪器融合方法。融合多尺度改进的跟踪算法和SuperDiMP跟踪算法,设计消失状态判断模块,综合考虑两种跟踪器的置信度分数及目标在预测框附近的可能性估计。实验结果表明,相较STARK跟踪算法,本文算法在GOT-10K数据集上的平均重叠率(AO)提升2.7%、成功率$SR_{0.5}$提高3.3%。在L-LaSOT数据集上,相较于STARK算法,成功率(AUC)提升0.8%,在目标消失重现挑战下成功率提升1%。

关　键　词:目标跟踪;多尺度窗口;Swin-Transformer;模板更新;多模型融合
中图分类号:TP391　　**文献标识码**:A　　**doi**:10.37188/CJLCD.2024-0118

Improved target tracking algorithm based on Swin-Transformer

LIU Shi, ZHU Ming*

(*Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun* 130033, *China*)

Abstract: An improved target tracking algorithm is proposed based on the Swin-Transformer network to address the problem of insufficient feature extraction capability and poor tracking effect often encountered when using convolutional neural networks in deep learning-based target tracking methods. Firstly, the window attention mechanism of the Swin-Transformer is enhanced across multiple scales, and a multi-scale window module termed MW-MSA is devised to extract more comprehensive local detail information. This augmentation, in conjunction with global contextual insights, engenders multi-scale discriminative features. Then, these features are integrated with the encoding-decoding structure of the Transformer, serving as the feature fusion network. An optimized multi-layer perceptron is employed as the update score judgment network to establish the state awareness module. Finally, a multi-tracker fusion method is proposed to address challenges like target occlusion and disappearance by integrating an improved tracking algorithm with the SuperDiMP tracking algorithm. Results from testing on L-LaSOT and GOT-10K datasets show significant improvements over the STARK tracking algorithm: a 2.7% increase in average overlap rate (AO) and a 3.3% increase in success rate (SR) on GOT-10K, and a 0.8% increase in success rate (AUC) on L-LaSOT. Moreover, under the target disappearance challenge, the success rate is improved by 1%.

收稿日期:2024-04-10;**修订日期**:2024-04-17.
*通信联系人,E-mail:zhu_mingca@163.com

Key words: target tracking; multi-scale windows; Swin-Transformer; template update; multi-tracker fusion

1 引 言

近年来,目标跟踪已经成为计算机视觉领域的一个重要研究方向,其应用范围涵盖了视频监控、智能交通、人机交互等多个领域。目标跟踪算法主要分为两大类:传统方法和基于深度学习的方法。传统方法[1-7]依赖于手工设计的特征[8-9]和经典的机器学习算法,虽然在某些场景下能够达到较好的跟踪效果,但普遍具有泛化能力差和适应性不强等问题。深度学习方法在视觉领域的快速发展为解决跟踪问题提供了新的视角和强大的工具。这些方法主要依赖于深度神经网络,特别是卷积神经网络[10](Convolutional Neural Network,CNN)来学习和提取视频序列中的复杂特征,从而实现对目标的精准跟踪。相比于传统方法,深度学习方法采用端到端的学习方式,无需手动设计特征,能够直接从大量标注数据中学习到前景和背景中更加深层次的信息,从而处理更复杂的场景变化和目标形态的变化,具有更好的泛化能力和鲁棒性。

然而,尽管基于深度学习的方法取得了显著的进步,但在实际应用中仍面临挑战。其中,主流的孪生架构中常采用卷积神经网络作为骨干网络,如ResNet[11]、AlexNet[12]等。虽然此种网络在图像分类任务中表现出色,但在目标跟踪任务中,这些网络的特征提取能力有限。特别是在处理目标遮挡、快速运动和背景干扰等复杂情况时,这些网络难以准确获取目标的关键特征。此外,卷积神经网络在处理图像时通常忽略了全局上下文信息,对前景与背景之间的长距离关系理解不足,导致跟踪过程中准确度下降。

Transformer[13]网络在计算机视觉领域的应用与发展为目标跟踪问题提供了新的解决方案。随着这种网络架构的逐步完善,基于Transformer的目标跟踪算法[14-19]逐渐受到关注并得到广泛研究。Transformer网络通过自注意力机制能够有效获取长距离关系,从而提高算法对复杂场景的适应能力。Swin-Transformer[20]通过动态窗口划分和层次化表示,能够更有效地获取全局上下文信息,从而提升模型对目标特征的理解和表达能力。

在此基础上,本文提出了一种基于Swin-Transformer改进的目标跟踪算法。在Swin-Transformer的基础上采用多尺度感受野,对注意力模块进行了多尺度的改进,优化网络的特征提取和表示能力,以应对复杂多变的跟踪场景。针对目标消失、快速移动的挑战,设计了多跟踪器融合方法。在多个标准数据集上的实验验证表明,该算法展现出优于现有深度学习方法的跟踪性能,证明了其在目标跟踪任务中的有效性。

2 Swin-Transformer网络结构

Transformer模型最初是在自然语言处理(NLP)领域提出的,并在这个领域取得了显著的成果。Vision-Transformer[21]首次尝试将Transformer引入到计算机视觉(CV)领域。随后,Swin-Transformer对模型进行了进一步的优化,引入了移位窗口等创新概念,使其成为计算机视觉领域一个重要的模型。Swin-Transformer的网络结构如图1所示。

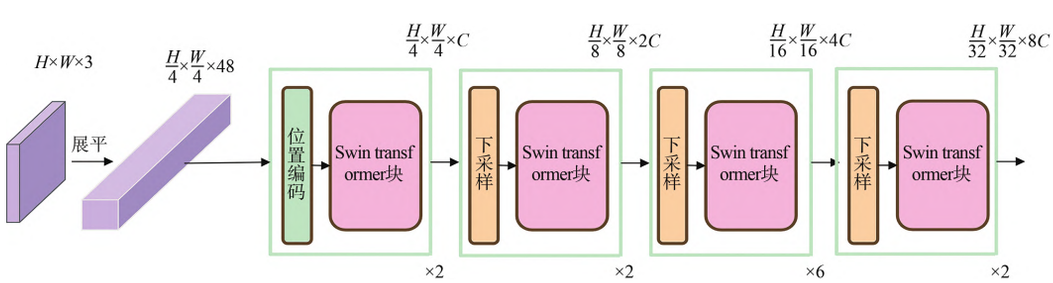


图1 Swin-Transformer网络结构图

Fig. 1 Diagram of Swin-Transformer network structure

首先，对图像进行分割，每4×4的相邻像素被划分为一个块。然后，对每块在通道方向上进行展平(Flatten)。在构建特征提取网络的过程中，通过4个不同的阶段(Stage)生成尺寸各异的特征图。对于第二阶段、第三阶段和第四阶段，每个阶段会降低来自前一阶段特征的尺寸进行下采样。在完成下采样后，特征图将传递至一系列的块(Block)结构中进行进一步的处理与特征提取。

其中，每一个块都由归一化层、窗口注意力模块、移位窗口注意力模块、多层感知机组成。特征图按照设置的$M\times M$的大小划分为单独的窗口，在每个窗口内部单独进行自注意力计算，该方法称为窗口注意力W-MSA。相较于原始的自注意力计算方式，此模块采用的方法减少了计算复杂度。结合注意力机制的计算方法的复杂度计算公式如公式(1)、(2)所示：

$$\Omega(\mathrm{MSA})=4hwC^2+2(hw)^2C\,, \tag{1}$$

$$\Omega(\mathrm{W\text{-}MSA})=4hwC^2+2M^2hwC\,, \tag{2}$$

其中：h代表特征图的高度，w为特征图的宽度，C代表特征图的深度，M为每个窗口的大小。

在采用窗口多头自注意力(W-MSA)模块的实现过程中，计算被限制在各自独立的窗口内，从而阻碍了不同窗口间的信息交流。为了解决这一问题，提出了移位窗口多头自注意力(SW-MSA)模块，通过对W-MSA窗口执行偏移操作，促进了窗口之间的信息互动。具体而言，将W-MSA模块中的窗口沿右下方向移动$M/2$个像素点，原有的M个窗口被扩增至$3M/2$窗口，这一改进使原始特征图中相邻窗口信息能够有效融合。在此过程中，原始图像中的分块区域移动到对角线方向，生成了一张包含新特征信息的特征图。

3　算法改进

针对Swin-Transformer网络对局部信息不敏感的问题，本研究对W-MSA进行改进，设计了多尺度窗口注意力模块MW-MSA。该模块不仅能够有效获取全局上下文信息，同时也可以注意到更多的细节信息，从而减小模型在跟踪过程中背景噪声导致的定位偏差的影响。随后，引入基于编码-解码器的特征融合模块，对提取到的特征进行进一步的筛选融合。优化多层感知机结构设计模板更新网络，结合特征融合网络构成状态判断模块，实现对模板信息的合理更新，以适应跟踪过程中外观变化、背景干扰等挑战。最后，针对目标消失、快速移动的挑战，设计多跟踪器融合方法。

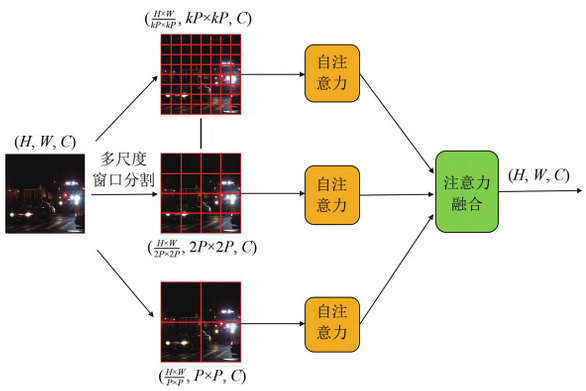


图2　多尺度窗口注意力MW-WSA示意图

Fig. 2　Diagram of multi scale window attention MW-WSA structure

3.1　多尺度窗口注意力

MW-MSA采用一种并行窗口结构，该结构由多个不同尺寸的窗口分支组成，每个分支覆盖了不同范围的注意力区域。输入特征图表示为$H\in R^{H\times W\times C}$，共有$k$个窗口分支，每个分支窗口大小为$\{P_i\}_{i=1}^{k}=\{P,2P,\cdots,kp\}$。特征图经过第$i$个分支窗口分割后，得到的尺寸变为$\left(\frac{H}{p_i}\times\frac{w}{p_i},p_i\times p_i\right)$。随后，注意力融合模块自适应地整合来自各个分支的信息。注意力的计算方法与Swin-Transformer一致，如式(3)所示，具体在每一个分支窗口中的计算公式如式(4)所示：

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\left(\frac{QK^{\mathrm{T}}}{\sqrt{d}}+\boldsymbol{B}\right)\boldsymbol{V}\right), \tag{3}$$

$$\hat{\boldsymbol{H}}_n^j=\mathrm{Attention}\left(\boldsymbol{H}^j\boldsymbol{W}_n^{Q},\boldsymbol{H}^j\boldsymbol{W}_n^{K},\boldsymbol{H}^j\boldsymbol{W}_n^{V}\right),\quad j=1,\cdots,\frac{HW}{(P_i)^2}\,, \tag{4}$$

其中：$\boldsymbol{W}_n^{Q}$、$\boldsymbol{W}_n^{K}$、$\boldsymbol{W}_n^{V}$分别代表$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$值，$\hat{\boldsymbol{H}}_n^j\in R^{P_i^2\times d_{h_i}}$。$d_{h_i}=\frac{C}{h_i}d$代表第$i$个分支中注意力

头的深度，其中 h_i 代表此分支中多头注意力中注意力头的个数。

首先，根据公式(5)、(6)计算单个分支中多头注意力的输出结果。其中 Y_n 为第 i 个分支中、多头注意力中第 n 个注意力头的输出结果，Y^i 为第 i 个分支中融合所有注意力头的输出结果。最后，经过注意力融合层，融合 k 个并行分支的结果之后，得到最终输出结果 Y，计算公式如式(7)所示。与 Swin-Transformer 不同，本文改进的多尺度窗口方法无需掩码、填充或者循环移位等数据操作，计算复杂度较低。具体的计算公式如式(8)所示。

$$\boldsymbol{Y}_n=\left[\hat{\boldsymbol{H}}_n^1,\hat{\boldsymbol{H}}_n^2,\cdots,\hat{\boldsymbol{H}}_n^{\frac{HW}{(P_i)^2}}\right],n=1,\cdots,h_i \tag{5}$$

$$\boldsymbol{Y}^i=\left[\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_{hi}\right] \tag{6}$$

$$\boldsymbol{Y}=\mathrm{FusionAttention}\left(\boldsymbol{Y}^1,\boldsymbol{Y}^2,\cdots\boldsymbol{Y}^k\right) \tag{7}$$

$$\mathcal{O}\left(\sum_{i=1}^{k}\frac{HW}{P_i^2}\times\frac{C}{hi}\times\left(P_i\times P_i\right)^2\times h_i+4\frac{HW}{P_i^2}\times P_i^2\times\left(\frac{C}{hi}\right)^2\times h_i\right). \tag{8}$$

采用 MW-MSA 组成块(Block)，其结构如图 3 所示。使用改进后的块替换原始的块结构，按照原始的 Swin-Transformer 网络架构组成新的骨干特征提取网络。其输入是一对图像：第一帧模板 $z\in R^{3\times H_z\times W_z}$ 和当前帧搜索区域 $z\in R^{3\times H_x\times W_z}$。经过骨干网络后，模板 z 和搜索图像 x 映射成为两个特征图 $f_z\in R^{C\times\frac{H_z}{s}\times\frac{H_z}{s}}$ 和 $f_x\in R^{C\times\frac{H_x}{s}\times\frac{H_x}{s}}$。

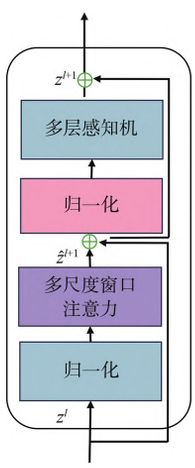


图 3 改进的 Swin-Transformer Block 结构图

Fig. 3 Structural diagram of improved Swin Transformer block

总的来看，在较大的窗口中，注意力机制能够捕捉更广泛的上下文信息，从而纠正较大的定位误差。而在较小的窗口中，注意力机制能在更为精细的范围内进行操作，以保留局部细节信息。采用多尺度多头注意力能够采用多尺度感受野，自适应地结合局部和全局上下文信息，从而提升跟踪算法的成功率和鲁棒性。

3.2 基于状态感知的模板更新模块

为了应对跟踪过程中的外观变化挑战，设计了状态感知模板更新模块。在跟踪的过程中，有很多种情况不适合进行模板更新，例如跟踪中常有背景相似物干扰、背景杂波等挑战，此时跟踪的预测框当中包含大量的背景信息，如果作为动态模板进行更新，将为后续的跟踪过程带来较大的误差。与此相反，如果仅采用第一帧模板信息作为固定模板，背景干扰、外观形变等因素可能会造成模板持续性污染，累积的误差会降低跟踪的鲁棒性。通过模板更新，跟踪器可以实时地学习到外观形变之后的特征，并且减少背景信息带来的干扰。所以对当前帧模板是否要进行更新的状态感知是十分重要的。

本文选取 Transformer 中的编码-解码结构作为特征融合网络对骨干网络输出的特征信息进行相似性判断，融合网络输出的信息中包含目标的位置、状态、大小以及时序信息。随后将其输入到经过批量归一化(Batch Normalization)层优化的多层感知机中，对当前状态进行判断，结构如图 4 所示，由一系列线性层和批量归一化层交替构成，实现正则化，以提升训练的稳定性和效率。最后一层输出前去掉激活函数，以便于输出当前状态的判断分数。

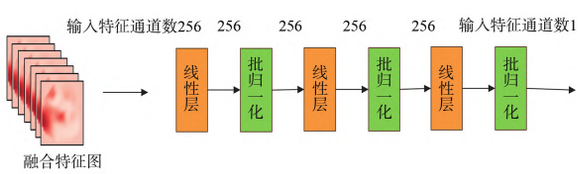


图 4 分数判断网络

Fig. 4 Score judgment network

加入状态感知模块后的整体网络结构图如图 5 所示。通过状态感知模板更新模块得到判断分数，判断是否需要进行模板更新，每 N 帧进行一次判断，如果分数大于阈值，并且预测框的位

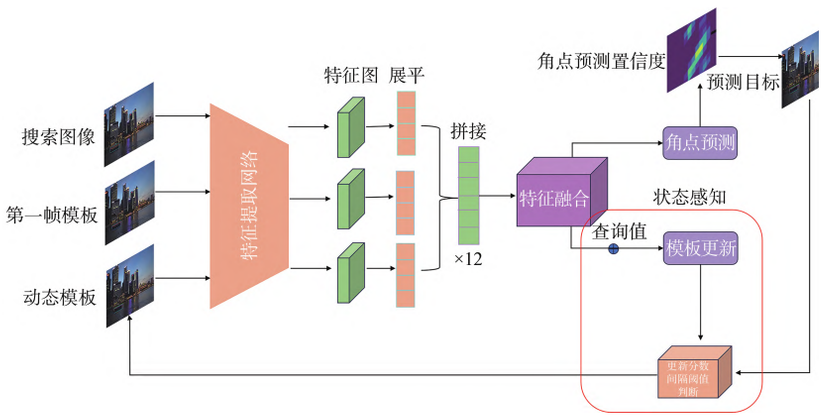


图5　引入模板更新后的跟踪网络结构图

Fig. 5　Diagram of tracking network structure after introducing template update

置在搜索区域内，则将当前帧判定为更新帧，并将当前帧跟踪预测头中预测的目标框提取，作为新一帧的动态模板。在跟踪过程中同时提取动态模板、第一帧模板、搜索区域的特征信息。

3.3　跟踪器消失状态判断模块

在跟踪过程目标面临消失、重现挑战时，STARK跟踪算法采用固定大小的搜索区域，无法有效应对上述挑战。为此，本文设计了多跟踪器融合的方法，采用SuperDiMP[15]跟踪器与上述多尺度改进后的跟踪器对目标进行联合跟踪。其中，改进后的跟踪器采用经过大量的数据集离线训练后得到的网络权重，从而提取到判别性更强的特征，并且改进了跟踪预测模块，得到的目标预测框更加准确。SuperDiMP跟踪器采用在线更新的训练方式，该算法在目标精度方面表现并不突出，但在处理诸如目标进出视野范围等情况时展现出较强的效能。因此，该跟踪算法在评估目标是否出现在画面中时的置信度评分具有更高的可信度。

在训练过程中，输入整段视频序列 $\mathcal{V}=\{F_t\in\mathcal{I}\}_{t=0}^{\mathrm{T}}$，视频共有 T 帧，其中每一帧序列记为 F_t，$\mathcal{I}=\{0,\cdots,255\}^{w\times h\times 3}$ 是序列中每一帧在RGB颜色空间的图片表示。将两个跟踪器分别定义为 τ^1 和 τ^2（其中SuperDiMP跟踪器为 τ^2），在第一帧中使用真值框位置信息对 b 进行初始化，为 b_0。在第 t 帧中，经过跟踪器预测后的预测框 b_t 使用左上角的坐标 (x_t,y_t)，预测框的宽度 w_t、高度 h_t 表示为 $b_t=[x_t,y_t,w_t,h_t]\in\mathcal{B}\subseteq\mathbb{R}^4$，采用置信度分数 $c_t\in[0,1]$ 表示当前序列目标存在的可能性。

参考基于跟踪、验证、检测的长时跟踪框架，在第 t 帧 F_t 中，将消失状态判断函数设置为 ν_t，该状态判断网络基于深度学习的二分类网络训练，训练数据为包含目标的候选框的正样本和不包含目标的候选框的负样本。正负样本的分类由候选样本和真值框之间的IOU值决定。在 b_t 附近，根据IOU的值生成多个候选框，将IOU值大于0.7的候选框设置为正样本，小于0.3的候选框设置为负样本。首先对网络进行预训练，随后通过正负样本对预训练的权重进行优化。

$v_t^{(i)}$ 表示在跟踪过程中，使用第 i 个跟踪器时，目标存在于第 t 帧视频序列预测框 b_t 附近的可能性。结合该预测框的置信度得分 $c_t^{(i)}$ 得到 $\hat{c}_t^{(i)}$，具体计算公式如式(9)所示。最终得到的 $\hat{c}_t^{(i)}$ 表示在第 t 帧序列中，目标仍存在于图像搜索区域的可能性，即为对目标消失状态的判断分数。

$$\hat{c}_t^{(i)}=\frac{c_t^{(i)}+\nu_t^{(i)}}{2}. \tag{9}$$

将 $\hat{c}_t^{(i)}$ 设置阈值为0.5。当 $\hat{c}_t^{(i)}$ 的值大于0.5时，目标判断为存在于视野区域，小于0.5则判断为消失，并将当前帧的目标状态记录为 $\hat{p}_t^{(i)}\in\{0,1\}$。但跟踪过程中的目标消失、重现是一个多变的动态过程，随着时间变化和误差累计，仅靠当前一帧图像的状态判断是不可靠的。针对此问题，结合当前帧之前 $\hat{T}$ 帧的状态信息，联合判断当前帧的目标存在状态。具体计算公式如式(10)所示，通常情况下将 $\hat{T}$ 的值设置为5。

$$\sum_{j=0}^{\hat{T}-1}\hat{p}_{t-j}^{(i)}>0.75\hat{T}. \tag{10}$$

如式(10)成立,则设置$p_t^{(i)}=1$,表示经过第i个跟踪器预测,当前帧目标仍存在于当前视野中;否则设置$p_t^{(i)}=0$,表示当前帧目标消失在视野当中。

通过上述的消失状态判断模块对当前帧的目标存在状态进行判断,并将当前状态进行记录。对比两种跟踪器判断的目标状态,如果两个跟踪器的判断结果相同且$p_t^{(i)}$的值都为1时,即通过两个跟踪器,目标都判断为存在于当前视野区域,未发生消失,此时τ^1的预测框精确度较高,所以选取τ^1的跟踪结果为最后的输出结果。如果两个跟踪器的状态判断结果只有一个为1,则采取结果判断为1的跟踪器的预测框为输出结果。如果两个跟踪器的判断状态值都为0,τ^1跟踪器相较于τ^2对丢失目标重新找回的能力更强,仍然采用τ^1的跟踪结果为最终输出结果。

4 实验与结果分析

4.1 数据集与评价指标

4.1.1 GOT-10K数据集

本文所采用的训练、测试数据集为GOT-10K[22],其覆盖范围广泛、类别丰富、注释比例大。该数据集提供了超过10 000个视频片段,包括超过150万个人工标记的目标框,能够实现跟踪器的统一训练和稳定评估。测试集包含420个视频,84类运动物体和31种运动形式,且每个类别最多取8个视频序列,以避免大规模的过拟合。验证集由训练集中随机抽样的180个视频构成。其中包含的跟踪挑战有遮挡、尺度变换、非刚性形变、非刚性旋转、光照变化、低分辨率等。

GOT-10K数据集中采用的度量指标为平均重叠度(Average Overlap,AO)和成功率(Success Rate,SR),以此验证网络模型的跟踪性能和整体效果,具体计算公式如式(11)、(12)所示。

$$\mathrm{AO}_i=\frac{\left|B_{\mathrm{g}_i}\cap B_{\mathrm{p}_i}\right|}{\left|B_{\mathrm{g}_i}\cup \mathrm{B}_{\mathrm{p}_i}\right|},\tag{11}$$

$$\mathrm{AO}=\frac{\sum_{1}^{n}\mathrm{AO}_i}{n},\tag{12}$$

其中:B_{g_i}为当前第i帧中真值边框,B_{p_i}为当前第i帧中预测边框,AO_i为两者的交集与并集的比值。将AO_i的阈值设置为0.5,则每一帧跟踪结果的AO_i大于0.5时判定为跟踪成功,否则判定为失败。将判定为成功的帧数除以总帧数得到跟踪成功率,表示为$\mathrm{SR}_{0.5}$。当阈值设置为0.75时,成功率表示为$\mathrm{SR}_{0.75}$。平均重叠率能够正确反映跟踪的准确性,值越大说明跟踪器的预测位置与真实位置越接近,跟踪效果越好。

4.1.2 LaSOT数据集

在复杂场景中,目标面临更加丰富多样的挑战。为了提高实验结果的准确性和可信度,本文选取LaSOT[23]数据集作为第二个数据集。LaSOT中所有序列都是长时的,平均视频长度超过2 500帧,共包含70种类别,每个类别由20个序列组成。其中包含的跟踪挑战有光照变化、完全遮挡、部分遮挡、非刚性旋转、运动模糊、快速运动、尺度变换、背景杂波、完全消失等。

LaSOT数据集中采用的评价指标为成功率(Success rate)和精确度(Precision)。首先对测试数据进行一次性评估(OPE),计算出不同跟踪算法的精确度和成功率。其中通过比较预测框C^{gt}和真值框中心点C^{tr}之间的欧式距离来计算算法的精确度。设置距离阈值为20,将距离高于阈值的跟踪器判定为跟踪失败。成功率的计算方法同GOT-10K数据集中相同,设置阈值的大小范围为0~1,绘制阈值与成功率曲线图,横坐标为判定成功设置的阈值,纵坐标为成功率。AUC(Area under Curve)为成功率曲线下面积,用于跟踪效果的展示。

LaSOT数据集中将目标消失在视野外、目标被完全遮挡的序列进行筛选,重新组建新的数据集L-LaSOT。该数据集共包含164个序列,其中目标消失的挑战共有99个视频序列,含有完全遮挡的序列共有65个视频。采用该数据集作为测试集。

4.2 实验配置

本文所有实验都是基于Ubuntu18.04系统、python 3.6版本、PyTorch深度学习框架完成,硬件配置为GeForce RTX™ 3090 GPU。选择GOT-10K、LaSOT训练集进行训练,搜索图片的大小为320×320,模板图像的大小为128×128。选取

AdamW 优化器作为网络优化器，衰减权重设置为 10^{-4}，主干网络和其余部分的初始学习率分别为 10^{-5}、10^{-4}，训练批次设置为 500。实验所用配置如表 1 所示。

表 1　实验配置

Tab. 1　Experimental configuration

Parameter	Configuration
OS	Ubuntu 18.04
Processor	Interl®Core(TM) i9-13 900KF
RAM	64.0 GB
GPU	NVIDIA Corporation Device 2 684
GPU Acceleration	CUDA 11.0

4.3　定性实验

为验证跟踪结果的有效性，在 L-LaSOT 数据集上选取 monkey9 序列做定性测试，结果如图 6 所示。其中红色预测框代表本章改进算法，蓝色框为 STARK_LT 跟踪算法，绿色框表示 TATrack 跟踪算法。Monkey9 序列为 L-LaSOT 数据集中的长时跟踪序列，共有 4 079 帧序列，包含了形变、快速移动和目标消失、重现等挑战。图 6 中选取其中含有目标消失、重现挑战的部分序列图片。从实验结果可以看出，在目标完全消失时，改进的跟踪器可以评估出此时的消失状态，并且不会跟踪到相似物上。STARK_LT 算法在前几帧并未判断准确，但经过几帧的判断后，在后续目标消失帧中判断成功。TATrack 算法则在目标消失状态跟踪到背景干扰物上，在多帧目标重现后继续跟踪目标。而融合算法由于对两个跟踪器联合判断，因此在目标消失时能够及时判断。总的来看，本文采用的融合多模型的跟踪算法在目标消失、重现的挑战上有很大的优势。

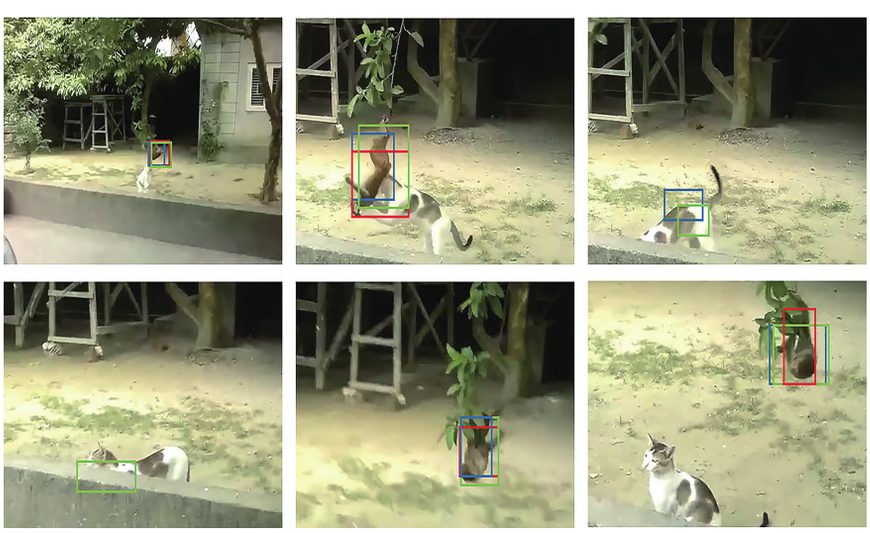


图 6　Monkey 长序列测试结果图

Fig. 6　Test results chart of Monkey long sequence

4.4　定量实验

4.4.1　数据集测试

在 L-LaSOT 数据集、GOT-10K 数据集上进行测试，结果如表 2、表 3 所示，其中(-SP)为不采用多模型融合之前的算法结果。对 L-LaSOT 数据集测试结果进行分析，本文算法的成功率较 STARK 算法提高 1%，但在精度上较 TATrack 算法有所下降。这可能是由于多尺度改进跟踪算法判断为目标消失时，选取 SuperDiMP 跟踪器进行跟踪，该算法的跟踪精度较 TATrack 算法较差。综合来看，在采用融合后的算法进行跟踪时，预测框 C^{gt} 和真值框中心点 C^{tr} 之间的欧式距离较大。根据表 2 和表 3 的结果分析可以看出，在 GOT-10K 和 L-LaSOT 数据集上，本文所提算法显著提高了跟踪评价指标。这一结果不仅体现了算法在理论上的合理性，也在实践中证明了其有效性。详细的数据对比和分析揭示了算法在各项指标上的具体改

表 2 L-LaSOT 数据集上的测试结果

Tab. 2 Test results on the L-LaSOT dataset %

算法	AUC	*P*
SiamRPN++[24]	40.6	40.8
SuperDiMP[15]	57.2	61.1
SiamR-CNN[25]	59.0	63.2
TransT[17]	57.8	62.4
TATrack[26]	61.1	69.2
STARK_LT[18]	61.6	67.5
Ours	62.4	67.8

表 3 GOT-10K 数据集上的测试结果

Tab. 3 Test results on the GOT-10K dataset %

算法	AO	$SR_{0.5}$	$SR_{0.75}$
SiamR-CNN	64.9	72.8	59.7
TrDiMP[14]	67.1	77.7	58.3
TransT	67.1	76.8	60.9
STARK	68.8	78.1	64.1
SBT-B[27]	69.9	80.4	63.6
Ours(-SP)	70.5	80.4	64.1
Ours	71.5	81.4	66.9

进幅度，进一步确认了其在复杂场景下的鲁棒性和适应性。

各算法在L-LaSOT数据集上的成功率如图7所示。可以看出本文算法相较于STARK_LT算法成功率提升0.8%。在目标消失在视野外、全遮挡、尺度变化、背景干扰等挑战下进行测试，结果如图8所示。结果显示，在目标消失在视野外的挑战中，本文算法的成功率较STARK_LT算

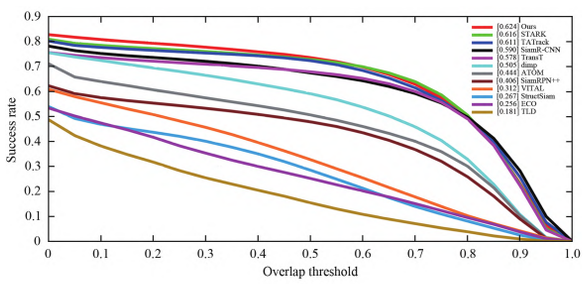
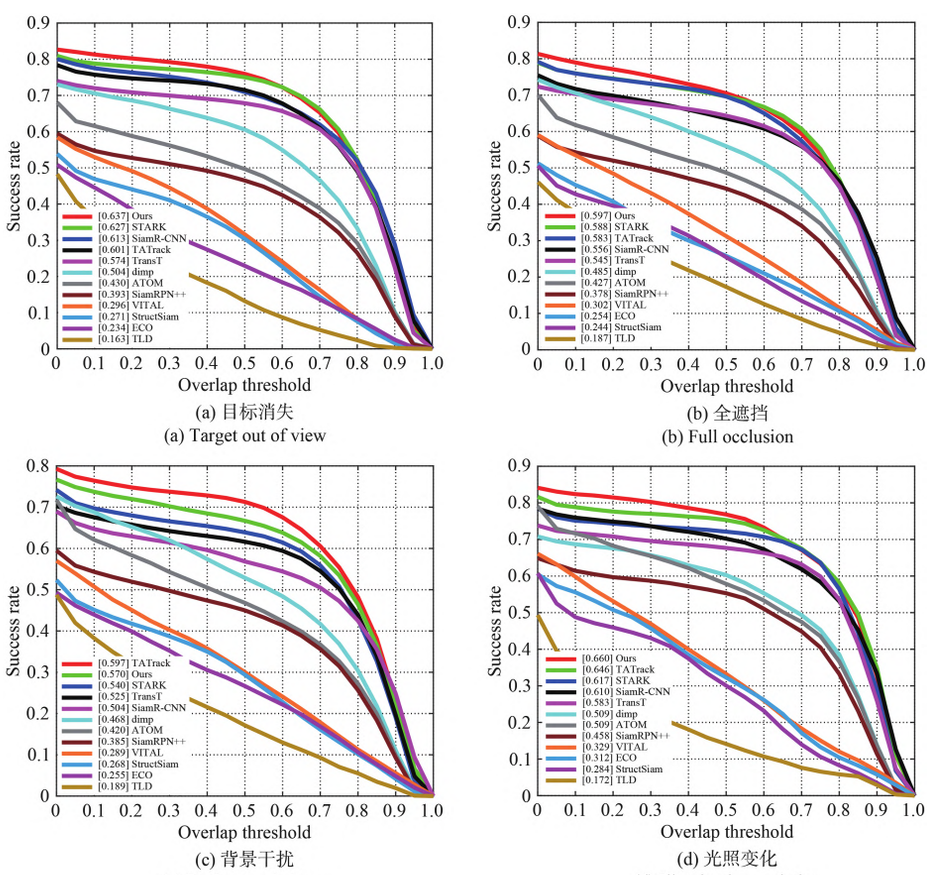


图 7 L-LaSOT 数据集上不同算法的成功率对比

Fig. 7 Comparison of the success rates of different algorithms on the L-LaSOT dataset



(a) 目标消失 (a) Target out of view

(b) 全遮挡 (b) Full occlusion

(c) 背景干扰 (c) Background clutter

(d) 光照变化 (d) Illumination variation

图 8 L-LaSOT 数据集上不同挑战下的成功率

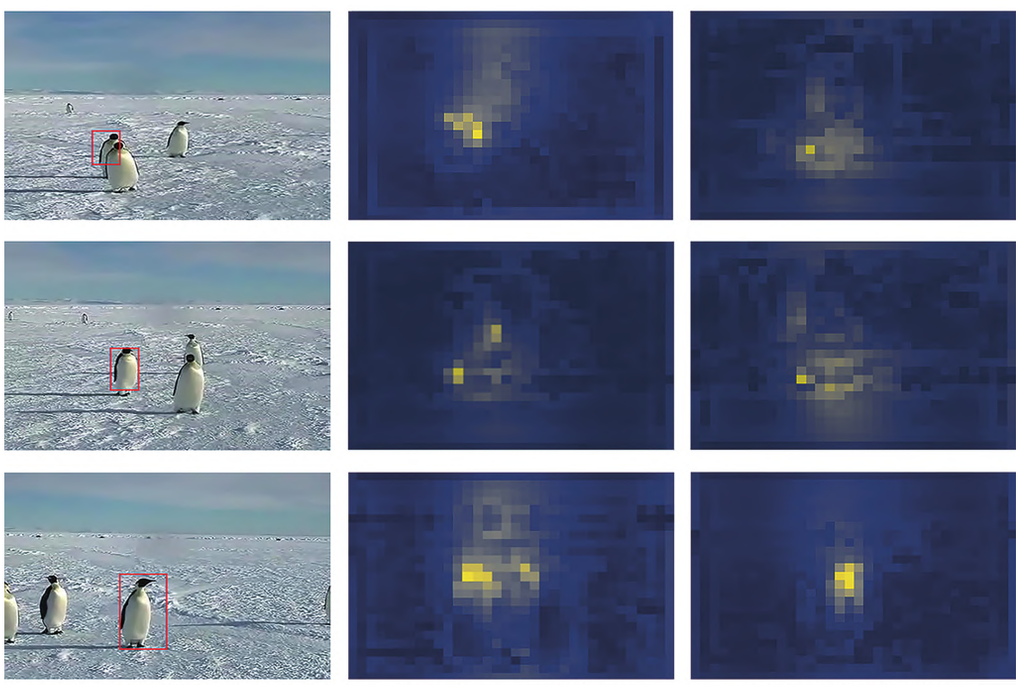
Fig. 8 Success rate under different challenges on the L-LaSOT dataset

法高1%,在全遮挡挑战中提升1.2%。但在背景干扰挑战中成功率相较于TATrack算法略低。

4.5　消融实验

4.5.1　固定模板

为了验证模板更新中固定模板(即第一帧模板)的必要性,在LaSOT数据集中选取序列,该序列中包含背景杂波干扰(主要为相似物干扰)等挑战,如图9所示,第二列得分图中跟踪算法采用本文设计的模板更新机制,即第一帧模板和动态模板共同作为模板信息。第一列跟踪预测置信度得分图中采用的算法仅采用动态模板但未加入第一帧模板。可以看出随着帧数的增加,背景干扰严重,在置信度得分图中有多个“亮点”,说明此时跟踪器跟踪到了多个相似性目标,造成跟踪成功率下降。分析其原因,是由于动态模板在更新过程中复杂的背景信息引入了过多的背景杂波,误差累积导致模板偏离真实目标,跟踪成功率降低。



(a) 原图 (a) Original images　(b) 未加入第一帧模板得分图 (b) Score maps (without the first frame template)　(c) 本章模板更新得分图 (c) Score maps (this chapter)

图9　不同更新方法置信度分数对比图

Fig. 9　Comparison chart of confidence scores for different update methods

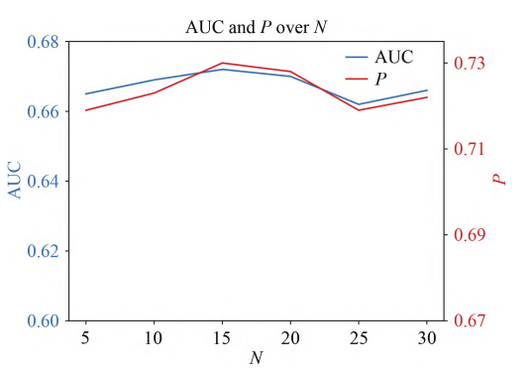


图10　不同间隔N值下在LaSOT数据集上的成功率和精度

Fig. 10　Success rate and accuracy on the LaSOT dataset with different interval N values

4.5.2　模板更新间隔

状态感知模块中的更新间隔N对跟踪结果也有很重要的影响。为了选取最为有效的间隔数,在LaSOT数据集上进行测试,计算不同N值下算法的成功率和精确度。由图10可以看出,当N值为15时,跟踪效果最好。

4.5.3　多跟踪器

为了验证跟踪器消失判断模块中两种跟踪器选取的有效性,对同一长时序列的前1 000帧的置信度分数以及IOU做记录,绘制于图11中。可以分析出,多尺度改进算法生成预测框的精度较高,但在目标置信度判断上能力较弱。而

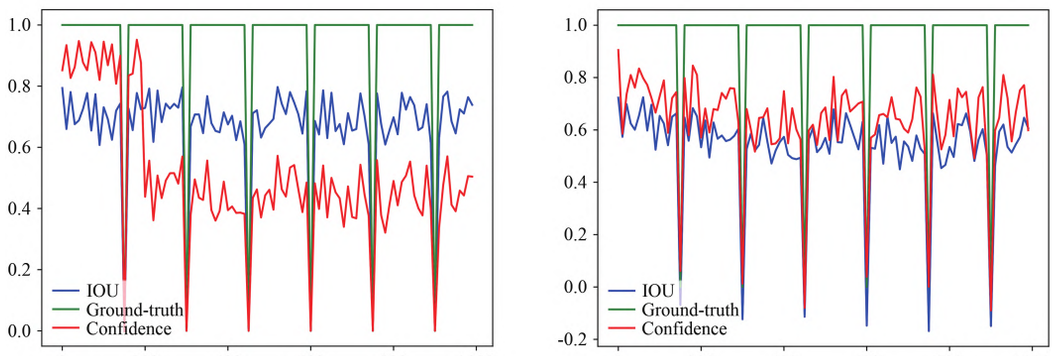


图 11 融合两算法在跟踪过程中的置信度和IOU得分

Fig. 11 Confidence and IOU scores of fusing the two algorithms in the tracking process

SuperDiMP算法则相反,生成的目标定位框的精度较低,但在置信度判断上更为准确。通过该实验证明了本文消失判断模块设置的合理性。

4.5.4 其他模块

从上述消融实验的结果中得出,每个组件均对实验成果产生了积极影响,如表4所示。具体来讲,采用ResNet-50作为骨干网络时,相比于多尺度改进算法,成功率降低了2.8%,这证实了特征提取网络在提升算法性能中的关键作用。在第二个实验中,替换骨干网络为Swin-Transformer后,相比于经过改进的网络,成功率降低了0.6%,进一步验证了多尺度模块在优化最终跟踪效果中的有效性。当模型移除编码部分时,成功率下降1.9%;而移除解码部分时,下降0.3%。这表明编码部分在实验中的影响更为显著。当省略模板更新模块时,成功率下降了2.1%,突出了模板更新在维持跟踪性能方面的重要性。最后,加入多模型融合模块,成功率提升1%。

表4 消融实验结果

Tab. 4 Results of ablation experiment %

Experiment	ResNet-50	Swin	MW-MSA	Enc	Dec	Score	SuperD	$SR_{0.5}$
1	√			√	√	√		77.6
2		√		√	√	√		79.8
3			√		√	√		78.5
4			√	√		√		80.1
5			√	√	√			78.3
6			√	√	√	√		80.4
7			√	√	√	√	√	81.4

5 结 论

本文提出的算法在L-LaSOT数据集上相较于STARK_LT算法,成功率(AUC)提升0.8%,在目标消失挑战下成功率提高1%。在GOT-10K数据集上相较于STARK算法,本文算法的平均重叠率(AO)提升2.7%,成功率$SR_{0.5}$提高3.3%。本文主要用多尺度优化的Swin-Transformer改进STARK跟踪算法骨干网络,并且优化状态感知的模板更新方法,最后采用多模型融合方法,针对目标消失、重现挑战做出优化设计。后续工作中将选取轻量型的网络模型,降低计算复杂度,以提升算法模型的实际应用性。

参　考　文　献:

[1] BOLME D S, BEVERIDGE J R, DRAPER B A, *et al*. Visual object tracking using adaptive correlation filters [C]. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010: 2544-2550.

[2] HENRIQUES J F, CASEIRO R, MARTINS P, *et al*. Exploiting the circulant structure of tracking-by-detection with kernels [C]. Computer Vision-ECCV 2012. Berlin, Heidelberg: Springer, 2012: 702-715.

[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] LEE J, LEE H, YI Y, *et al*. Making 802.11 DCF near-optimal: design, implementation, and evaluation [J]. *IEEE/ACM Transactions on Networking*, 2016, 24(3): 1745-1758.

[5] MCLEOD D R, GRIFFITHS R R, BIGELOW G E, *et al*. An automated version of the digit symbol substitution test (DSST) [J]. *Behavior Research Methods & Instrumentation*, 1982, 14(5): 463-466.

[6] POSSEGGER H, MAUTHNER T, BISCHOF H. In defense of color-based model-free tracking [C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Boston, MA, USA: IEEE, 2015: 2113-2120.

[7] DANELLJAN M, KHAN F S, FELSBERG M, *et al*. Adaptive color attributes for real-time visual tracking [C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA: IEEE, 2014: 1090-1097.

[8] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). San Diego, CA, USA: IEEE, 2005: 886-893.

[9] VAN DE WEIJER J, SCHMID C, VERBEEK J. Learning color names from real-world images [C]. 2007 IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA: IEEE, 2007: 1-8.

[10] YAMASHITA R, NISHIO M, DO R K G, *et al*. Convolutional neural networks: an overview and application in radiology [J]. *Insights into Imaging*, 2018, 9(4): 611-629.

[11] HE K M, ZHANG X Y, REN S Q, *et al*. Deep residual learning for image recognition [C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA: IEEE, 2016: 770-778.

[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] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need [C]//*Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach, CA, USA: Curran Associates Inc., 2017: 6000-6010.

[14] WANG N, ZHOU W G, WANG J, *et al*. Transformer meets tracker: exploiting temporal context for robust visual tracking [C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville, TN, USA: IEEE, 2021: 1571-1580.

[15] BHAT G, DANELLJAN M, VAN GOOL L, *et al*. Learning discriminative model prediction for tracking [C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South): IEEE, 2019: 6181-6190.

[16] YANG K, ZHANG H J, SHI J Y, *et al*. BANDT: a border-aware network with deformable transformers for visual tracking [J]. *IEEE Transactions on Consumer Electronics*, 2023, 69(3): 377-390.

[17] CHEN X, YAN B, ZHU J W, *et al*. Transformer tracking [C]//*Proceedings of* 2021 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Nashville, TN, USA: IEEE, 2021: 8122-8131.

[18] YAN B, PENG H W, FU J L, *et al*. Learning spatio-temporal transformer for visual tracking [C]//*Proceedings of* 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal, QC, Canada: IEEE, 2021: 10428-10437.

[19] MAYER C, DANELLJAN M, BHAT G, *et al*. Transforming model prediction for tracking [C]//*Proceedings of* 2022 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. New Orleans, LA, USA: IEEE, 2022: 8721-8730.

[20] LIU Z, LIN Y T, CAO Y, *et al*. Swin transformer: hierarchical vision transformer using shifted windows [C]// *Proceedings of* 2021 *IEEE/CVF International Conference on Computer Vision*. Montreal, QC, Canada: IEEE,

2021：9992-10002.

[21] DOSOVITSKIY A，BEYER L，KOLESNIKOV A，*et al*. An image is worth 16×16 words：transformers for image recognition at scale [C]. 9th International Conference on Learning Representations. Online：ICLR，2021.

[22] HUANG L H，ZHAO X，HUANG K Q. GOT-10k：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.

[23] FAN H，LIN L T，YANG F，*et al*. LaSOT：a high-quality benchmark for large-scale single object tracking [C]// *Proceedings* of 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach，CA，USA：IEEE，2019：5369-5378.

[24] LI B，WU W，WANG Q，*et al*. SiamRPN++：evolution of Siamese visual tracking with very deep networks [C]// *Proceedings of* 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Long Beach，CA，USA：IEEE，2019：4277-4286.

[25] VOIGTLAENDER P，LUITEN J，TORR P H S，*et al*. Siam R-CNN：visual tracking by re-detection [C]// *Proceedings of* 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Seattle，WA，USA：IEEE，2020：6577-6587.

[26] HE K J，ZHANG C L，XIE S，*et al*. Target-aware tracking with long-term context attention [C]//*Proceedings of the* 37*th AAAI Conference on Artificial Intelligence*. Washington，DC，USA：AAAI，2023：773-780.

[27] XIE F，WANG C Y，WANG G T，*et al*. Correlation-aware deep tracking [C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). New Orleans，LA，USA：IEEE，2022：8741-8750.

作者简介：

刘　时，女，硕士研究生，2021年于中国海洋大学获得学士学位，主要从事图像处理与视觉目标跟踪方面的研究。E-mail：liushi211@mails.ucas.ac.cn

朱　明，男，博士，研究员，主要从事视频图像处理、自动目标识别技术及成像目标跟踪等方面的研究。E-mail：zhu_mingca@163.com