激光与光电子学进展 先进成像

融合记忆信息的单目标跟踪模板更新机制

毛昱雯[1,2]，葛宝臻[1,2*]，权佳宁[1,2]，陈其博[1,2]
[1]天津大学精密仪器与光电子工程学院，天津 300073；
[2]天津大学光电信息技术教育部重点实验室，天津 300073

摘要 针对孪生架构单目标跟踪算法存在的目标状态更新不及时的问题，基于模板与记忆信息动态融合的跟踪策略，提出一种通用的模板更新机制。该机制采用双模块融合的更新策略：通过记忆融合模块融合搜索图像特征的短期记忆信息，获得目标变化情况；将前一帧可信的跟踪结果作为动态模板，从相关特征的角度，通过权重融合模块对原始模板和动态模板进行加权融合，通过结合跟踪过程的原始记忆与短期记忆实现更准确的目标定位。将模板更新机制应用于SiamRPN、SiamRPN++和RBO三种主流算法，并在VOT2019公开数据集上进行实验验证。结果表明：应用该机制后算法的性能得到了有效提升，具体而言，在SiamRPN++算法中，平均重叠期望值提升了6.67%，准确性提升了0.17%，鲁棒性下降了5.39%；此外，在遮挡、形变和背景干扰等复杂场景下，添加模板更新机制的SiamRPN++算法展现出较好的跟踪性能。

关键词 目标跟踪；孪生网络；模板更新；记忆信息
中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP231552

Template Update Mechanism for Single Target Tracking Incorporating Memory Information

Mao Yuwen[1,2], Ge Baozhen[1,2*], Quan Jianing[1,2], Chen Qibo[1,2]
[1]*School of Precision Instrument and Opto-Electronics Engineering, Tianjin University, Tianjin* 300073, *China*;
[2]*Key Laboratory of Opto-Electronics Information Technology, Ministry of Education, Tianjin University, Tianjin* 300073, *China*

Abstract Single target tracking algorithm based on Siamese architecture suffers from untimely target state update. To address this issue, a generic template update mechanism is proposed based on the dynamic fusion of templates and memory information. The mechanism uses a dual module fusion update strategy. The short-term memory information of search feature map is fused using a memory fusion module to capture target variations. The trusted tracking result of the previous frame is used as a dynamic template. The original and dynamic templates are fused using a weight fusion module from the correlated feature perspective to achieve more accurate target localization using the original and short-term memories during the tracking process. The template update mechanism is applied to three mainstream algorithms, SiamRPN, SiamRPN++ and RBO, and experiments are conducted on the VOT2019 public dataset. The results show that the performance of the algorithms is effectively improved after applying the mechanism. Specially, for the SiamRPN++ algorithm, the average overlap expectation is improved by 6.67%, the accuracy is improved by 0.17%, and the robustness is enhanced by 5.39% after applying the template update mechanism. In addition, the SiamRPN++ algorithm with the mechanism has better tracking performance in complex scenarios with occlusion, deformation and background interference.

Key words target tracking; Siamese network; template update; memory information

1 引 言

目标跟踪[1]是计算机视觉技术中的重要任务之一。计算机视觉技术通过对目标进行检测和跟踪，实现运动行为理解，可应用于视频监控[2]、自动驾驶[3]和无人机航拍[4]等领域。单目标跟踪是常见的目标跟踪任务，其实现方法可以分为传统的相关滤波方法和深度学习方法。相关滤波方法通过快速傅里叶变换将时

收稿日期：2023-06-19；**修回日期**：2023-07-11；**录用日期**：2023-07-24；**网络首发日期**：2023-08-18
基金项目：国家自然科学基金(61535008)
通信作者：*gebz@tju.edu.cn

域的相关运算转换到频域以加快跟踪速度[5]；深度学习方法通过神经网络得到准确的深度特征以提升跟踪精度。深度学习模型主要的网络结构分为孪生网络结构[6]和生成式对抗网络结构[7]。生成式对抗网络可以缓解训练样本分布不平衡的问题，但训练不易收敛；孪生网络的两个分支采用共享权重的卷积神经网络，可以减少网络参数量。同时，孪生的结构提供了额外的网络训练约束，训练更容易收敛。

孪生网络算法将第一帧的目标作为模板，在后续视频帧中选取搜索区域，通过学习模板和搜索区域的相似度匹配函数预测目标位置。其典型网络包括SiamRPN[8]、SiamRPN＋＋[9]和RBO[10]等。SiamRPN算法借鉴区域候选网络(RPN)[11]，通过锚框回归目标位置和形状，以局部单目标检测的方式提升算法的跟踪精度；SiamRPN＋＋算法将深度残差网络[12]引入孪生网络中，提高了特征提取网络容量；RBO算法针对跟踪定位和分类任务不平衡的问题，引入两个排名损失函数作为优化约束，提升了跟踪定位准确度。

这些方法取得了不错的效果，但跟踪时目标会遇到遮挡和形变等情况，使用单一固定的模板无法使跟踪网络获取目标的这些变化情况。因此，出现了动态更新模板的算法，典型的算法有GradNet[13]和STARK[14]算法。GradNet和STARK算法分别通过挖掘梯度信息和将时空信息与视觉变换网络(ViT)[15]集成来实现模板参数的在线更新。但基于ViT的STARK算法使用单尺度特征，会导致目标在发生尺度变化时，算法的跟踪精度下降。GradNet算法利用梯度信息不断更新模板以适应目标的变化，但这种方式会导致误差信息不断积累，准确的初始帧模板信息占比逐渐减小，进而增加跟踪失败的可能性。因此，本文提出一种既能保留初始帧的目标信息，又能反映目标状态变化的跟踪方法。

针对上述问题，本文提出融合记忆信息的模板更新机制(UM)，该机制采用基于搜索图像特征和相关特征的双模块融合策略，即：通过记忆融合模块对相邻帧之间的搜索图像特征进行加权融合，修正目标的变化情况；将可信的前一帧跟踪结果作为动态模板，从相关特征的角度，通过权重融合模块将第一帧的初始模板和动态模板融合，将初始模板的原始记忆与相邻帧间的短期记忆结合，以提升跟踪精度。将模板UM应用于SiamRPN、SiamRPN＋＋和RBO三种算法，在VOT2019[16]公开数据集上进行实验验证，实验结果表明，应用该机制后的算法性能均得到有效提升。此外，将所提算法与典型模板动态更新算法GradNet和STARK进行了比较，实验结果表明，所提算法取得了最好的效果。

2　算法框架和原理

2.1　模板UM

典型的孪生架构跟踪算法如图1所示。输入由模板图像z和待跟踪的搜索图像x构成，通过权重共享的深度神经网络(DNN)提取输入图像的特征，得到模板图像特征F_z和搜索图像特征F_x；进一步通过深度互相关操作[9]($\tilde{\mathrm{C}}$)计算得到两特征间的相似度。深度互相关操作以F_z为卷积核在F_x上逐通道进行卷积计算得到相关特征F_{corr}，令F_{corr}分别通过带有预测跟踪框坐标的回归头和判断预测跟踪框是否属于前景的分类头，选取前景得分最高的区域为跟踪的结果。该算法的特点是采用固定的初始帧目标实例作为模板图像，但仅采用初始帧作为模板图像，网络无法适应目标在跟踪中发生的形变和遮挡等变化。

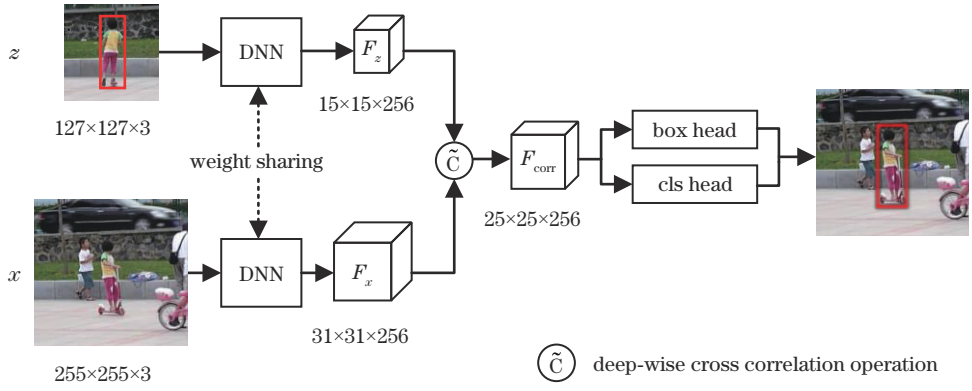

图1　孪生架构跟踪算法

Fig. 1　Tracking algorithm based on Siamese architecture

为了缓解固定模板无法反映图像动态变化的不利影响，提出融合记忆信息的模板UM，其基本结构如图2所示，其输入除初始模板z_1、搜索图像x之外，增加了动态模板z_i(i表示视频序列的第i帧图像)。此外，提出了一种双模块融合的更新策略，即：通过记忆融合模块(MFM)融合相邻视频帧的搜索图像特征，即：通过权重融合模块(WFM)融合初始模板和动态模板的相关特征。MFM将前两帧的搜索图像特征与当前帧的搜索图像特征进行加权融合，通过引入相邻帧间的短期记忆信息获知目标在帧间的变化情况。在WFM

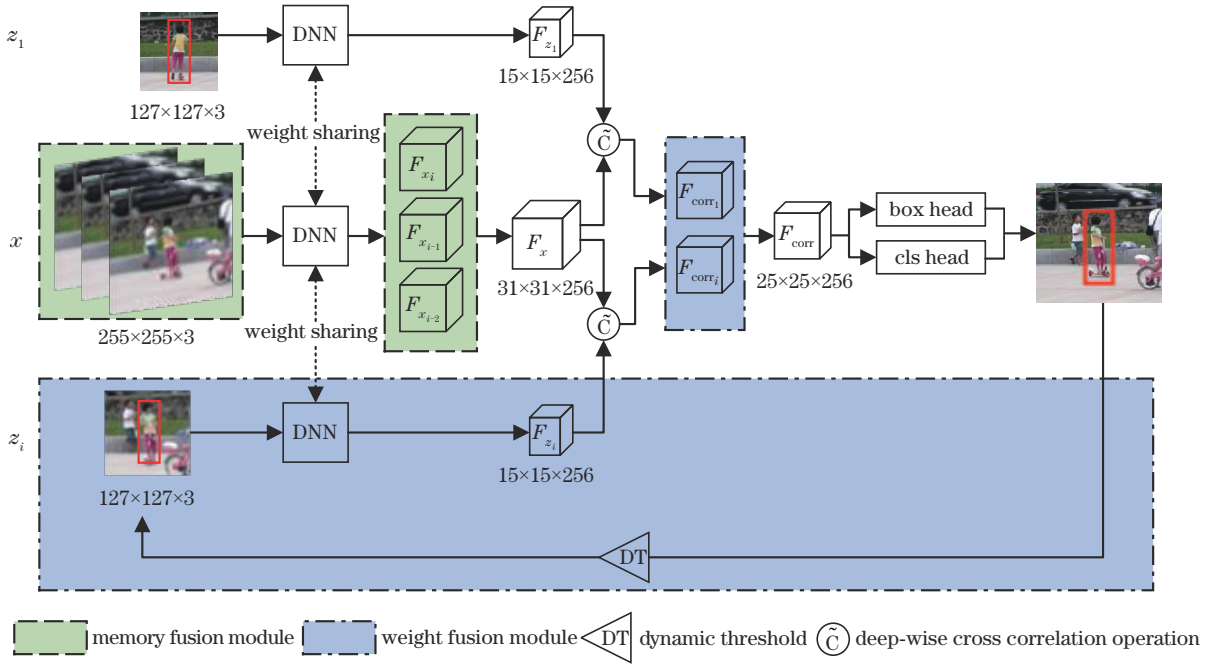

图 2　加入模板 UM 的跟踪算法

Fig. 2　Tracking algorithm with the template UM

中，采用动态阈值（DT）策略来判断前一帧跟踪结果是否准确。将准确的跟踪结果作为动态模板，可以使网络在目标发生遮挡等变化时通过动态模板的变化准确地跟踪目标。

2.2　MFM

为了增强网络在目标发生变化或受到遮挡等干扰情况时的跟踪能力，引入了短期记忆信息。通过 MFM 对相邻帧的图像特征进行加权融合，MFM 的结构如图 3 所示。首先获取 MFM 的融合系数，将初始模板特征 F_{z_1} 与当前帧搜索图像特征 F_{x_i} 经过互相关[17]后得到响应图 R_1。R_1 表示不同位置属于目标的得分高低，互相关通过卷积实现，表示为

$$R_1 = F_{z_1} * F_{x_i} + b, \tag{1}$$

式中：$*$ 表示以 F_{z_1} 作为卷积核在 F_{x_i} 上进行的卷积操作；b 表示模拟相似度的偏移量。将 R_1 的响应值通过

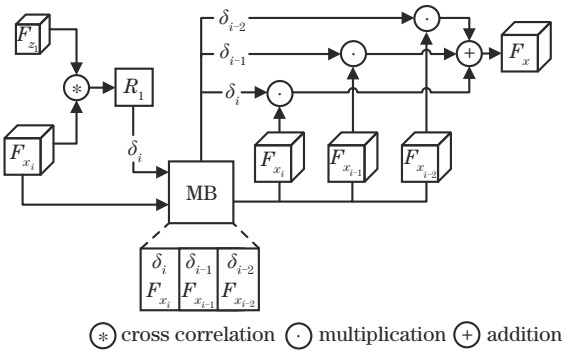

图 3　MFM 结构

Fig. 3　Structure of MFM

Softmax 函数进行可视化，可以得到 R_1 的得分概率分布图和概率最高区域的跟踪结果，如图 4 所示。从图 4(a) 中可以看出，R_1 表现出单峰分布，其响应值越大，表示

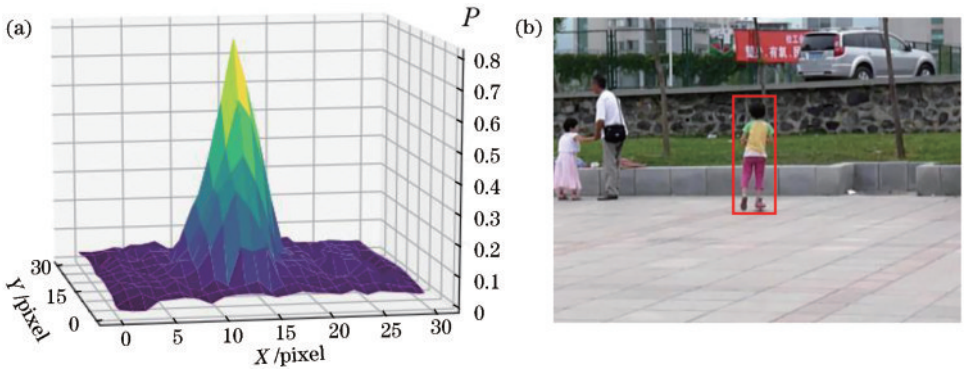

图 4　响应图的概率信息和对应的跟踪结果。(a)响应图的概率信息；(b)跟踪结果

Fig. 4　Probability of response map and the tracking result. (a) Probability of response map; (b) tracking result

该区域是跟踪目标的可能性越高，因此采用峰值作为融合系数。

接着引入前序的视频帧信息，在设定的记忆字典（MB）中存储前序搜索图像特征 $F_{x_{i-1}}$、$F_{x_{i-2}}$ 和融合系数 δ_{i-1}、δ_{i-2}，进一步与 F_{x_i} 和 δ_i 进行加权融合，表示为

$$F_x=\frac{\delta_i\cdot F_{x_i}+\delta_{i-1}\cdot F_{x_{i-1}}+\delta_{i-2}\cdot F_{x_{i-2}}}{\delta_i+\delta_{i-1}+\delta_{i-2}}。\tag{2}$$

具体实现上，MB 采用队列的形式，保存每一帧图像经过 MFM 后的搜索图像特征和融合系数。

2.3　WFM

在跟踪任务中，跟踪的目标会发生形变，被其他物体遮挡或被相似的物体或背景颜色干扰。如果更新的模板不准确，就会引入错误信息，从而导致算法的跟踪精度下降。因此，设计了一种 UM 来获取动态模板，通过 WFM 融合动态模板与初始模板。

通过 DT 判断是否对模板进行更新。选取响应图 R 的最大值 $r_{\max}$ 作为更新标准之一，记为 t_{th1}。由图 4 可知，t_{th1} 越大，该区域是目标的可能性越高。当 t_{th1} 满足 DT 的条件时，对模板进行更新。同时，将 R 的平均峰值相关能量（APCE）[18] 作为另一个更新标准，记为 t_{th2}，用来反映 R 的波动情况。其计算表达式为

$$t_{\text{th2}}=\frac{\left|r_{\max}-r_{\min}\right|^2}{\text{avg}\left[\sum_{x,y}\left(r_{x,y}-r_{\min}\right)^2\right]},\tag{3}$$

式中：$r_{\max}$ 和 $r_{\min}$ 分别为 R 的最大值和最小值；avg(·) 为平均值函数；$r_{x,y}$ 为 (x,y) 处的响应值。式（3）的分子反映了 R 的峰值大小，分母反映了 R 整体的波动状况。从上述分析可知，跟踪的目标没有被遮挡或干扰时，R 仅会产生单个峰值且波动较小。

DT 表示为

$$t_{\text{th1},i}\geqslant\mu\cdot\text{avg}\left(\sum_{i=0}^{N}t_{\text{th1},i}\right),\tag{4}$$

$$t_{\text{th2},i}\geqslant\mu\cdot\text{avg}\left(\sum_{i=0}^{N}t_{\text{th2},i}\right),\tag{5}$$

式中：μ 为权重值，用来控制模板更新的频率；i 为视频序列的第 i 帧图像，$i\in\{1,\cdots,N\}$，N 为视频总长度。当满足式（4）和式（5）的条件时，即当前帧的 t_{th1} 和 t_{th2} 以一定的权重大于历史帧的平均时，更新动态模板。这两个条件分别考虑了图像中的目标是否可信以及 R 的整体波动状况。

当同时满足上述更新条件时，将前一帧的跟踪结果作为动态模板 z_i。通过 DNN 提取 z_i 的特征，得到 F_{z_i}：

$$F_{z_i}=\text{DNN}(z_i),\tag{6}$$

式中，DNN(·) 为 DNN 操作。

将 MFM 获得的搜索图像特征 F_x 与 F_{z_i} 进行深度互相关（$\tilde{\text{C}}$）操作，得到动态模板的 F_{corr_i}：

$$F_{\text{corr}_i}=\tilde{\text{C}}\left(F_{z_i},F_x\right)。\tag{7}$$

将 F_x 与初始模板特征 F_{z_1} 进行深度互相关操作，得到初始模板的 F_{corr_1}：

$$F_{\text{corr}_1}=\tilde{\text{C}}\left(F_{z_1},F_x\right)。\tag{8}$$

将 F_{corr_1} 与 F_{corr_i} 通过图 5 所示的 WFM 进行融合，表示为

$$F_{\text{corr}}=\omega\cdot F_{\text{corr}_1}+(1-\omega)\cdot F_{\text{corr}_i},\tag{9}$$

式中，ω 为融合参数，$\omega\in[0,1]$。ω 越大，初始模板相关特征所占的比重越大。

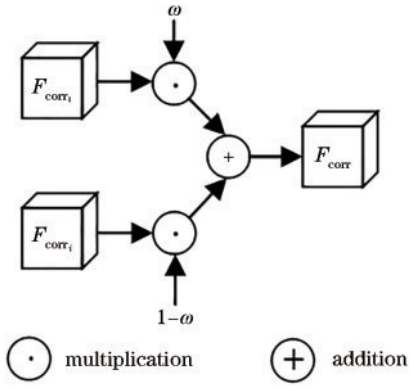

图 5　WFM 结构

Fig. 5　Structure of WFM

令 F_{corr} 分别经过由 1×1 卷积和激活函数组成的分类头和回归头，选取分类头得分最高时对应的跟踪框为跟踪结果。这时的跟踪结果兼顾了初始模板的原始记忆和动态模板的短期记忆，准确性更高。

2.4　损失函数

网络损失函数由分类损失和回归损失组成，表示为

$$L=\varepsilon_1\times L_{\text{cls}}+\varepsilon_2\times L_{\text{reg}},\tag{10}$$

式中：L_{cls} 为分类损失；L_{reg} 为回归损失；ε_1 和 ε_2 为超参数，取与 SiamRPN++ 中相同的值，分别为 1.0 和 1.2。

分类损失采用交叉熵函数，表示为

$$L_{\text{cls}}=-\left[n\times\ln(\hat{n})+(1-n)\times\ln(1-\hat{n})\right],\tag{11}$$

式中：n 为真实值；$\hat{n}$ 为跟踪器的预测结果。

回归损失采用 Smooth L1 函数，表示为

$$L_{\text{reg}}=\begin{cases}0.5d^2, & |d|<1\\ |d|-0.5, & |d|\geqslant1\end{cases},\tag{12}$$

式中，$d=j-\hat{j}$，表示目标的真实边界框与预测边界框的差值，其中，j 和 $\hat{j}$ 取边界框左上角和右下角的坐标的 4 个值。

3　实验与结果分析

3.1　实验方案

为了评估算法的性能，在 VOT2019[16] 数据集上进行测试。该数据集包含 60 个具有挑战性的视频序列，

是单目标跟踪领域中最常用的公开数据集之一。使用准确性(accuracy)、鲁棒性(robustness)、平均重叠期望(EAO)和跟踪速率[speed /(frame/s)]作为评价指标。

准确性的计算方法为

$$A(n)=\frac{1}{N_{\text{valid}}}\sum_{n=1}^{N_{\text{valid}}}\phi(n,k), \tag{13}$$

式中:$\phi(n,k)$为跟踪结果与真实标注框的交并比;N_{valid}为有效帧的数量。准确性是跟踪器在一段视频上重复运行k次的平均值,其数值越大,跟踪器的效果越好。

鲁棒性用来统计跟踪目标丢失的次数,计算方法为

$$R=\frac{1}{N}\sum_{i=1}^{N}F(i,k), \tag{14}$$

式中:$F(i,k)$为跟踪失败帧数;N为视频序列总帧数。鲁棒性是跟踪失败次数占总跟踪次数的比例,其数值越小,跟踪器的效果越好。

EAO的计算方法为

$$\varphi(n)=\frac{1}{N}\sum_{n=1:N}A(n), \tag{15}$$

$$E=\frac{1}{N_{\text{r}}-N_{\text{l}}}\sum_{n=N_{\text{l}}:N_{\text{r}}}\varphi(n), \tag{16}$$

式中:$\varphi(n)$是在以视频长度n为横坐标、准确性A为纵坐标的坐标系中,根据视频初始帧至不同长度视频帧下的准确性值绘制出的平均重叠曲线;N_{r}和N_{l}分别表示选取的面积占总曲线面积的50%时所对应的两个点的横坐标。将该范围内的平均准确性值作为EAO结果,能够更准确地反映出跟踪器的性能,EAO的数值越大,跟踪器的效果越好。

跟踪速率的计算方法为

$$S=N_{\text{s}}/T, \tag{17}$$

式中:N_{s}为跟踪到的视频帧数量;T为时间间隔。跟踪速率是每秒钟跟踪的视频帧数量,其数值越大,跟踪器的性能越好。

3.2 参数设置

算法通过PyTorch 1.8.1框架实现,并在两块NVIDIA GeForce RTX 2080Ti上训练,显存为11 GB,编程环境为Python 3.7。训练不同算法对比它们的性能时,训练参数与原算法保持一致。采用训练后的算法模型进行测试,训练数据集为COCO[19]、ImageNet[20]和YouTube-BB[21],将数据集中的图像经过裁剪等预处理操作输入模型中,其中模板图像尺寸为177 pixed×177 pixed,搜索图像尺寸为255 pixed×255 pixed,跟踪框选取的长宽比为0.33、0.50、1.00、2.00、3.00。将ImageNet数据集预训练的ResNet-50模型作为特征提取网络,共进行20个epoch的训练,前5个epoch采用warm up学习策略,后15个epoch以指数的形式衰减,学习率从0.001升为0.005后降为0.00005。在测试过程中,为了减轻实验随机误差的影响,选取5次测试结果平均值作为性能评价结果。通过VOT2019数据集上的实验确定式(4)、(5)中的参数μ和式(9)中的参数ω的取值,如表1、2所示。其中,评价指标旁边向上和向下的箭头符号分别表示该指标越大或越小,跟踪的性能越好。

从表1和表2中可以看出,当μ和ω的取值为0.75和0.90时,跟踪算法的EAO、准确性和鲁棒性达到最优,分别为0.304、0.592和0.456。由于模板更新过程对速率的影响不大,因此这种取值下的跟踪器具有最优的性能。

表1 参数μ对跟踪性能的影响

Table 1 Effects of parameter μ on tracking performance

Index	μ=0.65	μ=0.70	μ=0.75	μ=0.80	μ=0.85	μ=0.90
EAO ↑	0.292	0.303	0.304	0.300	0.299	0.297
Accuracy ↑	0.578	0.593	0.592	0.588	0.589	0.580
Robustness ↓	0.474	0.461	0.456	0.455	0.462	0.469
Speed /(frame/s) ↑	31	32	32	32	32	32

表2 参数ω对跟踪性能的影响

Table 2 Effects of parameter ω on tracking performance

Index	ω=0.75	ω=0.80	ω=0.85	ω=0.90	ω=0.95	ω=1.00
EAO ↑	0.295	0.297	0.302	0.304	0.300	0.295
Accuracy ↑	0.580	0.588	0.591	0.592	0.586	0.589
Robustness ↓	0.462	0.462	0.460	0.456	0.460	0.466
Speed /(frame/s) ↑	32	32	32	32	32	33

3.3 三种不同基准算法的消融实验

为了检验算法效果,在VOT2019数据集上对算法进行评估。选取主流目标跟踪算法SiamRPN[8]、SiamRPN++[9]和RBO[10]作为基准算法,并采用这三种算法以及加入模板UM的SiamRPN-UM、SiamRPN++-UM和RBO-UM算法进行消融实验,结果如表3所示。

由表3可知,对于三种基准算法,加入模板UM后算法的EAO、准确性和鲁棒性均有较好的结果。SiamRPN++-UM算法的性能提升最多,其中:EAO提升了6.67%,准确性提升了0.17%,鲁棒性降低了5.39%,但跟踪速率降低了8.57%。

3.4 主流目标跟踪算法的对比实验

选取了5种跟踪算法,包括SiamRPN、SiamRPN++、RBO、GradNet[13]和STARK[14]算法进行对比实验,与表3中性能较好的SiamRPN++-UM算法(记为Ours)进行对比,结果如表4所示。

由表4可知,在这6种算法中,Ours算法以最优EAO(0.304)、次优准确性(0.592)和最优鲁棒性(0.456),使跟踪器获得了较好的性能。虽然Ours算法的准确性低于STARK和RBO算法,但EAO和鲁

表 3 三种基准算法的消融实验结果

Table 3 Results of ablation experiments with 3 benchmark algorithms

Tracking algorithm	EAO ↑	Accuracy ↑	Robustness ↓	Speed /(frame/s) ↑
SiamRPN	0.247	0.566	0.592	160
SiamRPN-UM	0.257	0.572	0.572	145
SiamRPN++	0.285	0.591	0.482	35
SiamRPN++-UM	0.304	0.592	0.456	32
RBO	0.263	0.600	0.502	30
RBO-UM	0.277	0.609	0.487	28

表 4 主流跟踪算法的对比实验结果

Table 4 Comparative experimental results of mainstream tracking algorithms

Tracking algorithm	EAO ↑	Accuracy ↑	Robustness ↓	Speed /(frame/s) ↑
SiamRPN	0.247	0.566	0.592	160
GradNet	0.259	0.588	0.524	80
RBO	0.263	0.600	0.502	30
SiamRPN++	0.285	0.591	0.482	35
STARK	0.301	0.602	0.475	31
Ours	0.304	0.592	0.456	32

棒性评价指标最好，说明 Ours 算法虽然在跟踪目标的边界框精度上有所降低，但成功跟踪到目标的次数最多。Ours 算法的跟踪速率达到 32 frame/s，虽然低于 GradNet 等算法，但满足实时跟踪（25 frame/s）的要求，并优于 STARK 和 RBO 算法。

3.5 基于 VOT2019 数据集的可视化实验

为了直观地展示算法对于遮挡、形变和背景干扰场景下的目标跟踪效果，选取 VOT2019 数据集上具有复杂场景的 girl、gymnastics1、nature 和 tiger 视频序列进行实验。其中：girl 选取第 30、64、92 和 106 帧图像，以展示女孩目标在受到形变、遮挡和背景干扰影响时算法的跟踪效果；gymnastics1 选取第 54、75、89 和 96 帧图像，以展示运动员目标在受到形变和背景干扰影响时算法的跟踪效果；tiger 选取第 120、131、140 和 153 帧图像，以展示老虎模型头部目标在受到旋转和遮挡影响时算法的跟踪效果；nature 选取第 202、228、241 和 260 帧图像，以展示小鸟目标在受到其他鸟类和水草遮挡影响时算法的跟踪效果。将表现最好的 SiamRPN++ 作为基准算法，并将其与 SiamRPN++-UM 算法的跟踪结果进行比较，结果如表 5 和图 6 所示。

表 5 基于 VOT2019 数据集的不同视频序列的对比结果

Table 5 Comparison results of different video sequences based on the VOT2019 dataset

Video sequence	Algorithm	EAO ↑	Accuracy ↑	Robustness ↓	Speed /(frame/s) ↑
Girl	SiamRPN++	0.551	0.674	0.237	35.1
	SiamRPN++-UM	0.599	0.693	0.200	32.3
Gymnastics1	SiamRPN++	0.690	0.561	0	35.9
	SiamRPN++-UM	0.697	0.567	0	33.0
Tiger	SiamRPN++	0.740	0.699	0	36.8
	SiamRPN++-UM	0.752	0.714	0	33.6
Nature	SiamRPN++	0.634	0.614	0.300	34.4
	SiamRPN++-UM	0.654	0.625	0.200	31.5

表 5 为 4 个视频序列的数值对比结果，其中：第 1 列为视频序列的名称；每个视频序列对应着 2 行数据，第 1 行数据是 SiamRPN++ 算法的跟踪结果，第 2 行数据是加入模板 UM 后 SiamRPN++-UM 算法的跟踪结果；其余 4 列分别为 EAO、准确性、鲁棒性和跟踪速率结果。从表 5 中可以看出：加入模板 UM 后，对于这 4 个视频序列，算法的 EAO、准确性和鲁棒性均有较好的结果；算法对 girl 视频的跟踪性能提升最大，EAO 提升了 8.71%，准确性提升了 2.82%，鲁棒性下降了 15.61%，跟踪速率下降了 7.98%；对 gymnastics1 视频的跟踪性能提升较小，EAO 提升了 1.01%，准确性提升了 1.07%，鲁棒性没有变化，跟踪速率下降了 8.08%。这里鲁棒性为 0 的原因是基准算法成功跟踪到了该视频序列的所有目标，因此鲁棒性没有发生变化。

图 6 为 4 个视频序列的可视化结果。为了清晰地

图 6　VOT2019 数据集不同视频序列的可视化结果。(a1)～(a4) girl 原始图像；(b1)～(b4) girl 跟踪结果；(c1)～(c4) girl 跟踪结果的局部放大图像；(d1)～(d4) gymnastics1 原始图像；(e1)～(e4) gymnastics1 跟踪结果；(f1)～(f4) gymnastics1 跟踪结果的局部放大图像；(g1)～(g4) tiger 原始图像；(h1)～(h4) tiger 跟踪结果；(i1)～(i4) tiger 跟踪结果的局部放大图像；(j1)～(j4) nature 原始图像；(k1)～(k4) nature 跟踪结果；(l1)～(l4) nature 跟踪结果的局部放大图像

Fig. 6　Visualization results of different video sequences for the VOT2019 dataset. (a1)-(a4) Original images of girl; (b1)-(b4) tracking results of girl; (c1)-(c4) local zoomed images of girl's tracking results; (d1)-(d4) original images of gymnastics1; (e1)-(e4) tracking results of gymnastics1; (f1)-(f4) local zoomed images of gymnastics1's tracking results; (g1)-(g4) original images of tiger; (h1)-(h4) tracking results of tiger; (i1)-(i4) local zoom images of tiger's tracking results; (j1)-(j4) original images of nature; (k1)-(k4) tracking results of nature; (l1)-(l4) local zoomed images of nature's tracking results

展示图像的效果，在选取的每帧图像中展示了三张图像，分别为原始图像、跟踪结果图像以及放大目标区域后的局部图像；红色实线框代表真值，绿色点划线框代表 SiamRPN++-UM 算法的结果，蓝色虚线框代表 SiamRPN++算法的结果。通过直观地比较算法的结果框与真值框的重叠程度，可以证明不同场景下算法的跟踪效果。

对于遮挡场景：1) girl 视频序列的第 106 帧图像中，跟踪目标女孩被老人遮挡，基准算法的跟踪框不仅覆盖了女孩，还包含了老人；2) tiger 视频序列的第 131 和 140 帧图像中，老虎玩偶被树叶部分遮挡，基准算法的跟踪框偏离了目标并包含了树叶；3) nature 视频序列的第 228 和 241 帧图像中，小鸟被前方的水鸟部分遮挡，基准算法只跟踪到小鸟的部分身体且第 241 帧图像中包含了大量的水草。而加入模板 UM 的算法在这些帧图像中的跟踪框与真值框重叠面积较大，相较于基准算法能准确地跟踪目标。

对于形变场景，即 girl 视频序列的第 92 帧和 gymnastics1 视频序列的第 89 帧图像中，目标均发生了较大的形变，基准算法的跟踪框均与真值框有一定的偏移，而加入模板 UM 的算法的跟踪框基本与真值框重叠。

对于背景干扰场景，即 gymnastics1 视频序列的第 89 和 96 帧图像中，目标运动员受到了裁判员的干扰，基准算法的跟踪框包含了裁判员，特别是第 96 帧图像的跟踪框与真值框偏移较大，而加入模板 UM 的算法在这 2 帧图像中均能准确地跟踪到目标。

综上所述，应用所提模板 UM 的算法在遮挡、形变以及背景干扰场景下的跟踪效果更好。

4　结　　论

针对基于孪生架构的单目标跟踪算法在跟踪过程

中因目标发生遮挡或形变等变化时与固定模板的匹配度不高而导致的跟踪精度低的问题，提出了一种模板UM。该机制利用跟踪时的帧间记忆信息，通过结合不同时刻的图像特征来提升背景干扰等情况下的目标跟踪效果。在VOT2019公开数据集上的消融实验和对比实验表明：该机制具有通用性，适用于孪生架构的不同单目标跟踪算法；应用该机制的算法的EAO和准确性提升，鲁棒性降低，总体跟踪精度有效提升；虽然跟踪速率有所降低，但也满足实时跟踪的要求。可视化结果表明，在遮挡、形变和背景干扰的复杂场景下，应用模板UM的算法也能展现出良好的跟踪性能。

参 考 文 献

[1] 郭业才，曹佳露，韩莹莹，等. 基于光谱匹配降维和特征融合的高光谱目标跟踪[J]. 光学学报，2023，43(20)：2012002.
Guo Y C，Cao J L，Han Y Y，et al. Spectral matching dimensionality reduction and feature fusion for hyperspectral target tracking[J]. Acta Optica Sinica，2023，43(20)：2012002.

[2] 惠冠程，李开放，辛明，等. 基于视频行人重识别和时空特征融合的跟踪算法[J]. 激光与光电子学进展，2022，59(12)：1215004.
Hui G C，Li K F，Xin M，et al. Tracking algorithm based on video pedestrian recognition and spatio-temporal feature fusion[J]. Laser & Optoelectronics Progress，2022，59(12)：1215004.

[3] Liu M，Zhao F，Yin J L，et al. Reinforcement-tracking：an effective trajectory tracking and navigation method for autonomous urban driving[J]. IEEE Transactions on Intelligent Transportation Systems，2022，23(7)：6991-7007.

[4] 吴捷，马小虎. 基于特征融合与通道感知的无人机红外目标跟踪算法[J]. 激光与红外，2023，53(4)：626-632.
Wu J，Ma X H. UAV infrared target tracking algorithm based on feature fusion and channel awareness[J]. Laser & Infrared，2023，53(4)：626-632.

[5] 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，June 13-18，2010，San Francisco，CA，USA. New York：IEEE Press，2010：2544-2550.

[6] 张子烁，宋勇，杨昕，等. 基于动态特征注意模型的三分支网络目标跟踪[J]. 光学学报，2022，42(15)：1515001.
Zhang Z S，Song Y，Yang X，et al. Target tracking of three-branch network based on dynamic feature attention model[J]. Acta Optica Sinica，2022，42(15)：1515001.

[7] 王红涛，邓淼磊，赵文君，等. 基于深度学习的单目标跟踪算法综述[J]. 计算机系统应用，2022，31(5)：40-51.
Wang H T，Deng M L，Zhao W J，et al. Survey on single object tracking algorithms based on deep learning[J]. Computer Systems and Applications，2022，31(5)：40-51.

[8] Li B，Yan J J，Wu W，et al. High performance visual tracking with Siamese region proposal network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition，June 18-23，2018，Salt Lake City，UT，USA. New York：IEEE Press，2018：8971-8980.

[9] Li B，Wu W，Wang Q，et al. SiamRPN++：evolution of Siamese visual tracking with very deep networks[EB/OL]. (2018-12-31)[2023-03-02]. https://arxiv. org/abs/1812.11703.

[10] Tang F，Ling Q. Ranking-based Siamese visual tracking[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)，June 18-24，2022，New Orleans，LA，USA. New York：IEEE Press，2022：8731-8740.

[11] Ren S Q，He K M，Girshick R，et al. Faster R-CNN：towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，39(6)：1137-1149.

[12] 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)，June 27-30，2016，Las Vegas，NV，USA. New York：IEEE Press，2016：770-778.

[13] Li P，Chen B，Ouyang W，et al. Gradnet：gradient-guided network for visual object tracking[EB/OL]. (2019-09-15)[2023-02-03]. https://arxiv.org/abs/1909.06800.

[14] Yan B，Peng H W，Fu J L，et al. Learning spatio-temporal transformer for visual tracking[EB/OL]. (2021-03-31)[2023-02-03]. https://arxiv.org/abs/2103.17154.

[15] Dosovitskiy A，Beyer L，Kolesnikov A，et al. An image is worth 16×16 words：transformers for image recognition at scale[EB/OL]. (2020-10-22)[2023-02-03]. https://arxiv.org/abs/2010.11929.

[16] Kristan M，Matas J，Leonardis A，et al. The seventh visual object tracking VOT 2019 challenge results[C]//2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)，October 27-28，2019，Seoul，Republic of Korea. New York：IEEE Press，2020：2206-2241.

[17] Bertinetto L，Valmadre J，Henriques J F，et al. Fully-convolutional Siamese networks for object tracking[M]//Hua G，Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham：Springer，2016，9914：850-865.

[18] Wang M M，Liu Y，Huang Z Y. Large margin object tracking with circulant feature maps[EB/OL]. (2017-03-15)[2023-02-03]. https://arxiv.org/abs/1703.05020.

[19] Lin T Y，Maire M，Belongie S，et al. Microsoft COCO：common objects in context[EB/OL]. (2014-05-01)[2023-02-03]. https://arxiv.org/abs/1405.0312.

[20] Russakovsky O，Deng J，Su H，et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision，2015，115(3)：211-252.

[21] Real E，Shlens J，Mazzocchi S，et al. YouTube-BoundingBoxes：a large high-precision human-annotated data set for object detection in video[EB/OL]. (2017-02-02)[2023-02-03]. https://arxiv.org/abs/1702.00824.