智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

时空融合与判别力增强的孪生网络目标跟踪方法

黄昱程, 肖子旺, 武丹凤, 艾斯卡尔·艾木都拉

引用本文:
黄昱程, 肖子旺, 武丹凤, 艾斯卡尔·艾木都拉. 时空融合与判别力增强的孪生网络目标跟踪方法[J]. 智能系统学报, 2024, 19(5): 1218-1227.
HUANG Yucheng, XIAO Ziwang, WU Danfeng, et al. Spatiotemporal fusion and discriminative augmentation for improved Siamese tracking[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(5): 1218-1227.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202306005

DOI: 10.11992/tis.202306005
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240829.1643.006

时空融合与判别力增强的孪生网络目标跟踪方法

黄昱程[1], 肖子旺[1], 武丹凤[2], 艾斯卡尔·艾木都拉[1]
（1. 新疆大学 计算机科学与技术学院, 新疆 乌鲁木齐 830046; 2. 北京联合大学 机器人学院, 北京 100101）

摘　要: 孪生跟踪器的出现极大提升了跟踪任务性能。然而, 当前跟踪器难以精准描述目标外观变化, 造成面临遮挡和尺度变化等挑战时的性能衰减。另外, 杂乱背景会产生干扰响应图, 误导目标定位。为此, 引入 2 个基于 Transformer 的跟踪模块用于提高孪生跟踪器性能。其中时空融合模块使用交叉注意力机制的全局特征关联, 迭代累积历史线索从而提高目标外貌变化的鲁棒性。判别力增强模块关联目标和搜索区域的语义信息, 以提高目标判别能力。此外, 使用空间通道加权特征融合, 充分发掘空间分布和语义相似性的时空信息。所提模块可嵌入主流孪生跟踪器, 在公开数据集上的实验证明了方案的优越性。

关键词: 人工智能; 深度学习; 计算机视觉; 目标跟踪; 神经网络; Transformer; 特征融合; 时序建模
中图分类号: TP18　　**文献标志码**: A　　**文章编号**: 1673−4785(2024)05−1218−10

中文引用格式: 黄昱程, 肖子旺, 武丹凤, 等. 时空融合与判别力增强的孪生网络目标跟踪方法 [J]. 智能系统学报, 2024, 19(5): 1218–1227.
英文引用格式: HUANG Yucheng, XIAO Ziwang, WU Danfeng, et al. Spatiotemporal fusion and discriminative augmentation for improved Siamese tracking[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1218–1227.

Spatiotemporal fusion and discriminative augmentation for improved Siamese tracking

HUANG Yucheng[1], XIAO Ziwang[1], WU Danfeng[2], HAMDULLA A[1]
(1. School of Computer Science and Technology, Xinjiang University, Urumqi 830046, China; 2. College of Robotics, Beijing Union University, Beijing 100101, China)

Abstract: The development of Siamese trackers has considerably enhanced the tracking performance. However, current trackers have difficulty accurately describing changes in the appearance of the target, which results in performance degradation under occlusion and scale changes. Cluttered backgrounds can interfere with the tracker response and mislead target localization. Therefore, two Transformer-based modules are introduced to improve the performance of Siamese trackers. Specifically, the spatiotemporal fusion module uses a cross attention mechanism for global feature association to iteratively accumulate historical clues for improving the robustness of the target appearance change. Meanwhile, the discriminative enhancement module associates semantic information between the target and the search area to enhance the target discrimination capability. In addition, adaptive weighted channel-spatial fusion is utilized to fully explore the spatiotemporal information of spatial distribution and semantic similarity. The proposed module can be embedded into mainstream Siamese trackers and exhibits superior performance on public datasets.

Keywords: artificial intelligence; deep learning; computer vision; object tracking; neural network; Transformer; feature fusion; temporal modeling

目标跟踪[1-2]是计算机视觉领域的一个基本任务。在无人机、自动驾驶和视频监控等领域有着广泛的应用。基于孪生网络的跟踪算法[3-8]是目前主流跟踪算法。将目标搜索区域图像和目标模板图像输入到共享权重参数的深度神经网络用于特征提取, 通过互相关运算匹配二者特征, 生

收稿日期: 2023−06−02.　**网络出版日期**: 2024−08−30.
通信作者: 艾斯卡尔·艾木都拉. E-mail: askar@xju.edu.cn.

成用于定位目标位置的特征响应图，实现跟踪。近年来，孪生网络跟踪器取得优越性能，使其成为众多研究的焦点。Bertinetto 等[5] 的开创性工作参考了传统的判别相关滤波 (discriminative correlation filter, DCF) 跟踪器[9-11]，并使用卷积神经网络代替了手工特征提取的过程，通过卷积操作实现了互相关。随着 Faster R-CNN(faster region-based convolutional neural network)[12] 在目标检测领域兴起，Li 等[6] 首次将其中的区域建议网络 (region proposal network) 引入到跟踪领域，提出了多分支预测，将跟踪变为二分类问题，利用回归方法精确地预测目标在当前帧的位置。Zhang 等[13] 首次尝试设计深度孪生神经网络应用于目标跟踪算法，并取得可观成绩。Li 等[14] 通过采用移位的图像增强方式来适应更复杂的网络结构，由此收集到丰富的语义信息，从而提高了互相关精确度。Guo 等[15] 采取无锚点进一步提高跟踪器的性能。

当前跟踪方案没有利用跟踪过程中的时空连续性，从而无法结合上下文推断出更为精确的目标定位。为了解决该问题，相关跟踪算法[16-17] 采用线性的更新方案，从而平滑地调整目标匹配模板。然而，在复杂场景下，光照变化、目标形变和局部遮挡等问题不断出现，因此需要构造一个自适应的目标描述器用于缓解外貌变化带来的性能衰减。对此，一些研究尝试利用自监督、光流和时空正则化项来构造出自适应的时空建模方法，提高跟踪器应对目标外貌改变的鲁棒性，但是仍难以关联全局时序信息[18-20]。背景杂乱同样是孪生网络跟踪面对的挑战性难题，其容易造成生成响应图中含有多个高响应语义描述点，使得跟踪器无法从语义空间判断目标和干扰的区别，最终造成跟踪失败。对此，Zhu 等[16] 通过在训练样本上构建语义干扰的负样本集，以数据驱动的方式提高模型判别能力。Wang 等[17] 和 Lukezic 等[21] 设计可学习的权重分配机制，使描述目标语义信息的通道比重更大，从而提高应对背景杂乱的鲁棒性。Danelljan 等[22] 直接利用空间正则化，通过限制杂乱背景区域，降低模型被干扰概率。然而，这些方法受到样本构造真实性的限制，同时没有建立起深度特征空间中各元素的语义关联性，难以充分利用深度特征空间中目标和干扰之间的语义差异。

受 Zhang 等[18] 工作的启发，本文通过利用 Transformer[23] 的自注意力机制来增强孪生网络跟踪器。所提方案缓解现有孪生网络跟踪器在利用时空信息和杂乱背景方面的局限性。为提高跟踪器对目标外观变化的鲁棒性，引入一个便携式的时空融合模块，该模块关联独立各帧，并且在各帧间传递丰富时空信息，利用自注意力机制的全局特征关联，实现对目标外貌变化的全局时空建模。为提高跟踪器在杂乱背景中区分目标的能力，提出一个判别力增强模块，该模块首先增强搜索区域特征的内部关联，进一步利用交叉注意力机制发掘目标模板和搜索区域之间的语义相似性。

1　改进工作

本节详细阐述本文设计的整体跟踪框架，如图 1 所示，主要包括 2 个关键部分：时空融合 (spaito-temporal fusion, ST) 模块以及判别力增强 (discriminative augmentation, DA) 模块。

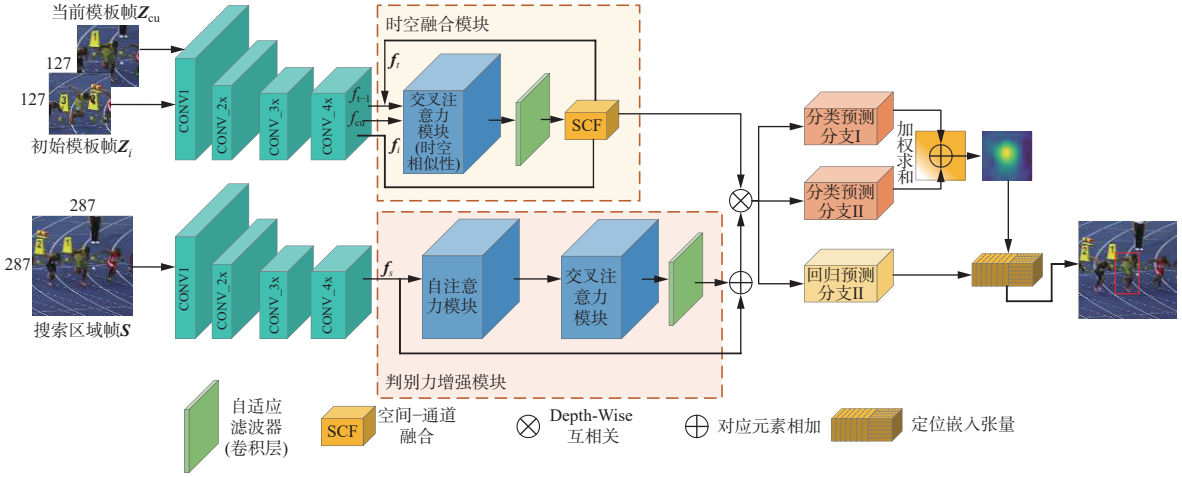


图 1　DASTSiam 整体框架

Fig. 1　Overall architecture of DASTSiam

1.1　时空融合模块

典型的孪生网络在整个跟踪过程中使用固定的初始模板进行匹配。具体地，初始模板特征 $\boldsymbol{Z}_i \in \mathbf{R}^{C\times N_z\times N_z}$ 和搜索区域特征 $\boldsymbol{X} \in \mathbf{R}^{C\times N_z\times N_z}$ 通过骨干

网络$\Psi(\cdot)$分别得到对应的深度特征，随后使用互相关操作生成含有定位信息的响应图：

$$\text{resMap} = \Psi(\boldsymbol{Z}_i) \times \Psi(\boldsymbol{X})$$

搜索区域特征$\boldsymbol{X}$通常会不断变化，其中包含的目标外貌随着时间进行改变，最终导致匹配失败。受到 UpdateNet 和 Transformer 启发，提出时空融合模块掘跟踪序列中的时空相似性，从而得到一个基于历史信息的概率分布用于自适应调整匹配模板。ST 模块结构如图 2 所示。ST 一共有 2 个输入，分别是上一时刻累积模板特征$\boldsymbol{f}_{t-1}$和当前跟踪器预测输出的模板特征$\boldsymbol{f}_{\text{cu}}$。ST 通过使用视频帧中的变化迭代$\boldsymbol{f}_{t-1}$使其不断获得过去目标状态并且更好地表示目标的外貌。同时，ST 通过$\boldsymbol{f}_{\text{cu}}$可以不断整合之前的目标状态到累积模板$\boldsymbol{f}_{t-1}$，从而生成新的累积模板特征$\boldsymbol{f}_{t-1}$。ST 模块定义为

$$\text{ST}(\boldsymbol{f}_{t-1}, \boldsymbol{f}_{\text{cu}}) = \Phi_{\text{ST}}\left(\boldsymbol{f}_{\text{softmax}}\left(\frac{W^q(\boldsymbol{f}_{\text{cu}})W^k(\boldsymbol{f}_{t-1})}{\sqrt{d}}\right)W^v(\boldsymbol{f}_{\text{cu}})\right)$$

式中：$W^i(\cdot)_{(i=q,k,v)}$是维度置换和线性映射的组合操作。首先，通过将$\boldsymbol{f}_{j(j=q,k,v)}$的维度从$C \times \frac{N_z}{2 \times S_{\text{stride}}} \times \frac{N_z}{2 \times S_{\text{stride}}}$变换到$\frac{N_z}{2 \times S_{\text{stride}}} \times \frac{N_z}{2 \times S_{\text{stride}}} \times C$，接着使用 3 个权重可学习的全连接层实现线性映射，即$W^q(\cdot)$、$W^k(\cdot)$、$W^v(\cdot)$，d是维度的大小。

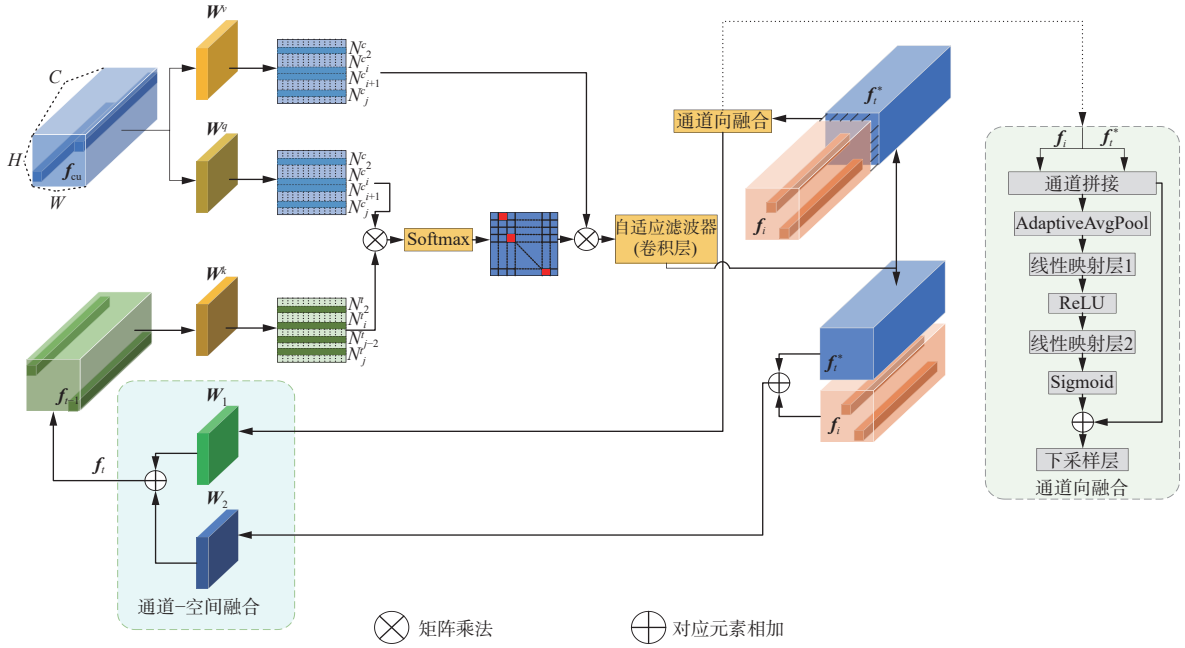


图 2　时空融合模块

Fig. 2　Spatio-temporal fusion module

通过维度转换和线性映射的组合操作以后，$W^q(\boldsymbol{f}_{\text{cu}})$和$W^k(\boldsymbol{f}_{t-1})$中的每一行代表了一个特征点向量，为了方便，分别用$\boldsymbol{N}_i^c \in \mathbf{R}^{\frac{N_z}{2 \times S_{\text{stride}}}}$和$\boldsymbol{N}_i^t \in \mathbf{R}^{\frac{N_z}{2 \times S_{\text{stride}}}}$对它们进行表示。随后进行矩阵乘法获得一个包含语义相似性分布的注意力矩阵$\boldsymbol{A}$，其定义为

$$\boldsymbol{A} = P(T_t | T_i, T_{i+1}, \cdots, T_{t-1}) = \sum_{i=0}^{N_z^*}\sum_{i=0}^{N_z^*} N_i^c N_i^{t\text{T}} \left(N_z^* = \frac{N_z}{2 \times S_{\text{stride}}}\right)$$

式中$\{T_i, T_{i+1}, \cdots, T_{t-1}\}$是历史模板。$\boldsymbol{A}$通过迭代融合每一步的当前匹配模板，从而充分利用时空上下文增强匹配模板的特征。由于$\boldsymbol{f}_t$具有广泛且可靠的历史信息，使得$\boldsymbol{f}_{\text{cu}}$能够更加关注特征空间中与历史目标具有时空相似的位置。

匹配模板能够通过时空上下文进行自适应调整，从而生成更合适的目标特征。然而，连续 2 帧图像发生巨大变化时，会发生历史信息误导，导致结果严重漂移。因此，通过 2 步进行校准。第 1 步是将 ST 的初步融合结果通过一个可学习的自适应滤波$\Phi_{\text{ST}}(\cdot)$进行调整。第 2 步，根据初始匹配模板具有真实的语义特征，从空间和通道 2 个方面与初始模板$\boldsymbol{f}_i$实现自适应的加权特征融合，更好地利用时空信息并提高跟踪器的鲁棒性。空间和通道的自适应加权融合公式为

$$\boldsymbol{f}_t = W_1(\Phi_{\text{ST}}(\boldsymbol{f}_{\text{cu}}, \boldsymbol{f}_{t-1}) + \boldsymbol{f}_i) + W_2(\Gamma(\Phi_{\text{ST}}(\boldsymbol{f}_{\text{cu}}, \boldsymbol{f}_{t-1}), \boldsymbol{f}_i))$$

式中：“+”代表对应元素相加，$\Gamma(\cdot)$是通道向融合。图 3 详细描述了融合过程。特征点向量的空间分布和每个特征点向量中的通道信息，在与初始模板特征进行融合时都至关重要。因此，通过 2 个可学习权重自适应调整融合结果。在此基础上，当前自适应生成的累积模板特征$\boldsymbol{f}_t$，将会通过迭代融合$\boldsymbol{f}_{\text{cu}}$从而提高跟踪器对时空信息的利用。

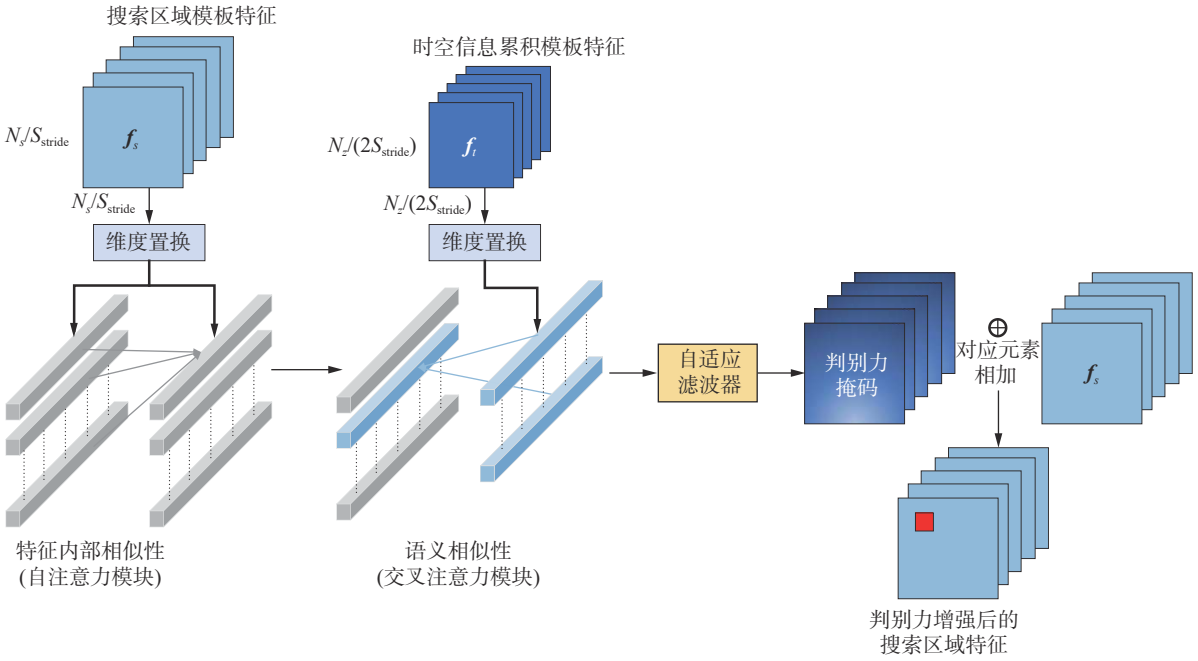


图 3　判别力增强模块

Fig. 3　Discriminative augmentation module

1.2　判别力增强模块

在通过 ST 得到最终的目标模板特征 $\boldsymbol{f}_t$ 以后，为了提高跟踪器的判别能力，使用 DA(模块结构如图 3 所示) 来增强搜索区域特征 $\boldsymbol{f}_s \in \mathbf{R}^{C\times\frac{N_s}{S_{\text{stride}}}\times\frac{N_s}{S_{\text{stride}}}}$。DA 模块的定义为

$$\boldsymbol{f}'_s = f_{\text{softmax}}\left(\frac{W^q(\boldsymbol{f}_s)W^k(\boldsymbol{f}_s)}{\sqrt{d}}\right)W^v(\boldsymbol{f}_s)$$

$$\boldsymbol{f}^*_s = \Phi_{\text{DE}}\left(f_{\text{softmax}}\left(\frac{W^q(\boldsymbol{f}'_s)W^k(\boldsymbol{f}_t)}{\sqrt{d}}\right)W^v(\boldsymbol{f}_t)\right)+\boldsymbol{f}_s$$

在 DA 中，首先使用自注意力机制增强 $\boldsymbol{f}_s$ 的内部特征元素之间的关联性，随后使用交叉注意力机制将 $\boldsymbol{f}_s$ 与 $\boldsymbol{f}_t$ 在特征空间中进行语义关联，从而生成一个与 $\boldsymbol{f}_s$ 大小一致的判别掩码，用来区分 $\boldsymbol{f}_s$ 中目标与干扰。与时空融合模块类似，采用一个可学习的滤波器用于抑制掩码特征空间中的干扰信息。最后用相加的方式将掩码作用于搜索区域特征，从而提高目标与干扰在特征空间中的差异。如图 4 所示，利用梯度热力图可视化，可以看出所提方案对目标外貌变化和杂乱背景的较强鲁棒性。

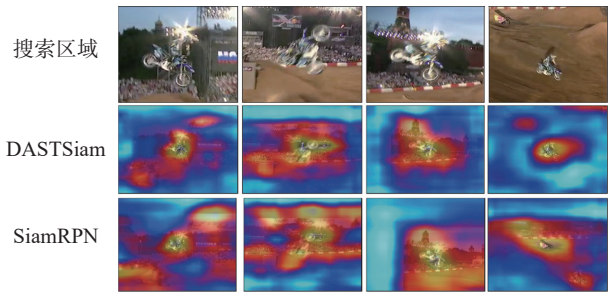


图 4　梯度热力图可视化

Fig. 4　Gradient heatmap visualization

1.3　训练策略

1.3.1　训练阶段

通过图像模糊、拼接以及增加噪点的方式设置了特殊样本，用于模拟复杂环境下的干扰。首先按规定从视频序列选取 3 帧进行裁剪得到 3 个匹配模板，用 $\boldsymbol{T}_i$、$\boldsymbol{T}_{t-1}$、$\boldsymbol{T}_{\text{cu}}$ 表示。在特征提取以后，分别得到 3 个特征，分别为时空累积特征 $\boldsymbol{f}_{t-1}$，前一帧预测出的目标特征 $\boldsymbol{f}_{\text{cu}}$ 和初始模板特征 $\boldsymbol{f}_i$。对于 $\boldsymbol{T}_i$，从连续的 50 个帧序列中随机选取，对于 $\boldsymbol{T}_{\text{cu}}$，从经过处理的特殊样本中选取，从而提高跟踪器对噪点和模板腐蚀的抗干扰能力。同时，随机选取连续 2 帧图像作为 $\boldsymbol{T}_{t-1}$ 和 $\boldsymbol{T}_{\text{cu}}$，充分训练时空融合模板利用帧间时空上下文的能力，在最大程度上拟合推理阶段。

1.3.2　损失函数

对于回归函数，直接使用与 SiamRPN 一致的平滑 L1 损失函数 (smooth-L1) 预测锚点中心到真实目标中心点的归一化距离。使用 A_x、A_y、A_w、A_h 表示锚点的中心坐标和宽高。使用 G_x、G_y、G_w、G_h 表示对应的真实样本的中心坐标和宽高。归一化误差公式为

$$\boldsymbol{\delta}[0] = \frac{G_x - A_x}{A_w}, \qquad \boldsymbol{\delta}[1] = \frac{G_y - A_y}{A_h}$$

$$\boldsymbol{\delta}[2] = \ln\frac{G_w}{A_w}, \qquad \boldsymbol{\delta}[3] = \ln\frac{G_h}{A_h}$$

式中 $\boldsymbol{\delta}$ 为误差张量。对于分类损失，根据文献 [24] 中的分析，在样本分配策略上，基于中心距离的正负样本分配方式，相比于交并比 (intersection

over union, IOU) 的方式，会带来更好均值平均精度 (mean average precision, mAP)。因此采用 Tian 等[25]基于中心距离的正负样本分配策略，公式定义为

$$c_{\text{pos}} = \Theta(C_{\text{pos}}) = \frac{C_{\text{pos}} - O_{\text{ori}}}{S_{\text{stride}}}$$

$$d^{*}_{x_i,y_i} = \left(\frac{c_{y_{\text{lt}}} + c_{y_{\text{rb}}}}{h} - r_i\right)^2 + \left(\frac{c_{y_{\text{lt}}} + c_{y_{\text{rb}}}}{w} - l_i\right)^2$$

式中：C_{pos}是边框（bounding box）的对角坐标，分别包括了左上角坐标 (left top, lt) 和右下角坐标 (right bottom, rb)。通过使用函数$\Theta(\cdot)$，将初始的边框坐标映射到分类分支输出的特征空间，得到对应的坐标c_{pos}。其中，O_{ori}是锚点 (anchor) 的初始值，r_i代表了特征图中的第i行，l_i代表了特征图中第i列。将$d^{*}_{x_i,y_i}$小于阈值的样本设置为正样本，否则设置为负样本。同时参考 FCOS，另外增加了一个二分类交叉熵 (binary cross entropy, BCE) 损失函数用于提高自适应更新模板的置信度。最终完整的损失函数公式为

$$L_{\text{cls}} = \lambda L_{\text{cls1}} + L_{\text{cls2}}$$

$$L_{\text{total}} = \lambda_1 L_{\text{cls}} + L_{\text{reg}}$$

式中：λ和λ_1是调整损失函数比重的超参数，L_{cls1}和L_{cls2}分别对应使用的 2 个损失函数。

1.3.3　推理阶段

初始阶段，直接采取初始帧$\boldsymbol{T}_i$的特征图作为$\boldsymbol{f}_i$、$\boldsymbol{f}_{t-1}$、$\boldsymbol{f}_{\text{cu}}$，其中$\boldsymbol{f}_i$是在初始的帧进行框定的真实目标边框。后续每一迭代，通过 ST 不断自适应生成累积模板特征$\boldsymbol{f}_t$。在最终预测阶段，通过加权计算分类分支的预测结果选取最高置信度的索引确定最终回归结果，生成高置信的目标边框，裁剪出当前模板$\boldsymbol{T}_{\text{cu}}$，通过骨干特征提取得到$\boldsymbol{f}_{\text{cu}}$，进一步配合$\boldsymbol{f}_t$实现下一次的时空信息的迭代融合。

2　实验与结果分析

2.1　实现细节

2.1.1　方法细节

改进工作在骨干网络为 ResNet50[26] 以及单层互相关的 SiamRPN 基础上，采用基于中心距离的正负样本设置方案，同时嵌入所提时空融合模块以及判别力增强模块，形成 DASTSiam 跟踪器。ResNet50 作为骨干网络，首先在 ImageNet[27] 上进行预训练，使用训练好的模型进行跟踪算法中的特征提取任务。本文跟踪算法的时空融合模块与判别力增强模块同是基于 Transformer 实现，Transformer 中编码器−解码器的输入张量维度为 256，查询 (query)、键 (key) 和值 (value) 的线性映射输出维度为 1 024，编码器和解码器各设置 4 层，注意力头数为 4，dropout 设置为 0.1。判别力模块中直接采取解码器部分执行交叉注意力生成判别力掩码。时空融合模块中采取 Transformer 编码器部分执行交叉注意力生成时空相似性矩阵，同时包含 1 个通道融合网络，网络的输入维度为 512，其中通道注意力的输入维度为 1，隐藏层维度为 15，最后通过1×1卷积层，将通道注意力加权后的特征张量通过卷积将维度由 512 映射为 256，完成多帧的通道融合。在推理阶段，每次只有 1 帧被送入网络执行特征提取。第 1 帧为匹配模板，后续每帧为跟踪器预测出的当前模板。在第 1 帧处理结束后，初始目模板的特征以及累积历史信息的模板特征都被放置在内存。

2.1.2　训练细节

训练数据由 LaSOT[28] 和 VID[29] 中划分出来的训练集组成。DASTSiam 跟踪器训练 80 个周期。骨干网络前面 15 个周期参数固定，并使用预热操作来调整 ST 和 DA 从而加速收敛。剩余的学习周期中，学习率采用对数级衰减，从 0.005 逐渐减小到 0.000 5。另外，仅训练骨干网络的最后 3 层。采用 SGD 作为优化器，batch size 为 64。模板和搜索图像的大小为 127 像素×127 像素，匹配模板的大小为 287 像素×287 像素。

2.1.3　实验平台

跟踪器的实现环境采用 Python3.6 以及 PyTorch1.1.0. 模型在 2 张 NVIDIA 24 GB 3090 GPUs 的服务器上训练。并在单张 NVIDIA 12 GB 3080Ti GPU 上进行测试。

2.2　测试基准和指标

为全面衡量所提方案，采用了 4 个被广泛使用的测量基准数据集 (OTB100[30]、VOT2018[31]、GOT10k[32] 和 LaSOT[28]) 来评估算法性能同时与其他最先进的跟踪器进行比较。

2.2.1　OTB100 测试基准

该数据集选取来自 98 个视频的 100 个测试集。主要的评估方案为成功率和精确度。成功率通过测量估计位置和真实目标的交并比来衡量，并通过曲线下面积 (area under the curve, AUC) 来与其他跟踪器进行比较。精确度则是首先计算预测目标和真实目标中心点距离的误差，并统计出误差小于设定阈值视频帧的百分比，由此来衡量不同跟踪器的性能。

2.2.2　VOT2018 测试基准

该数据集通过短时跟踪序列来测试跟踪器性能。VOT2018 使用 3 个不同的指标来进行性能测

试。首先是正确性 (accuracy, A), 主要衡量预测的边框与真实边框的交并比。其次是鲁棒性 (robustness, R) 衡量跟踪过程中目标丢失的次数。最后是平均期望重叠率 (expected average overlap, EAO), 通过关联 A 和 R 2 个指标来综合衡量跟踪器的性能。

2.2.3　GOT10k 测试基准

该数据集是包含了 563 个目标类别, 包含了大量的野外移动目标。为了进行训练和测试, 该数据集被划分为训练集、测试集以及验证集。其包含了 2 个评估指标。首先是平均重叠率 (average overlap, AO), 其通过测量目标所有结果的平均重复率来衡量跟踪器在目标尺度和位置预测上的正确性。其次是成功率, 其衡量每个测试序列中重复率到达一定阈值时所占帧的百分比, 主要为 50%(SR_{50}) 和 75%(SR_{75})2 个阈值。

2.2.4　LaSOT 测试基准

该数据集是一个大范围的长期单目标跟踪数据集, 包括许多高质量的人工标注。该数据集共包含 1 400 个视频序列, 共有 70 个类别, 每个类别包含 20 个子序列。其标注包括边框、自然语言描述等。与 OTB100 类似, LaSOT 采用一次性评估手段 (one-pass evaluation, OPE)。其设置了 14 个具有挑战的测试属性: 光照变化 (illumination variation, IV)、全遮挡 (full occlusion, FOC)、局部遮挡 (partial occlusion, POC)、形变 (deformation, DEF)、动态模糊 (motion blur, MB)、快速移动 (fast motion, FM)、尺度变化 (scale variation, SV)、旋转 (rotation, ROT)、长宽比变化 (aspect ratio change, ARC) 等。LaSOT 使用 3 个指标: 精确度、归一化精确度、成功率来评价跟踪器在具体环境下的性能。

2.3　消融实验

实验通过控制数据集、训练策略、平台配置等变量一致, 同时减少无关干扰的影响, 从而营造公平的评估和验证环境, 实现有效的消融分析。

2.3.1　时空融合模块消融分析

为了验证该模块的有效性, 采取 SiamFC 和 Modified SiamRPN 作为基线。保持训练策略一致。如表 1 所示, 通过在 OTB100 上的多次实验, 当 ST 模块嵌入基线模型后性能得到提升。因为 ST 模块充分发掘帧间关系从而利用历史信息增强当前模板特征, 同时结合 Transformer 注意力机制对特征内部的全局关联特性, 实现有效的全局时空建模。SiamFC 的成功率提高了 1.6%, Modified SiamRPN 的成功率提高了 4.3%。

表 1　ST 模块消融分析

Table 1　Ablation of ST module

跟踪器	ST	OTB100	
		成功率/%	精度
SiamFC		58.32	0.77
	√	**59.30**	**0.78**
Modified SiamRPN		60.70	0.80
	√	**63.34**	**0.83**

2.3.2　判别力增强模块消融分析

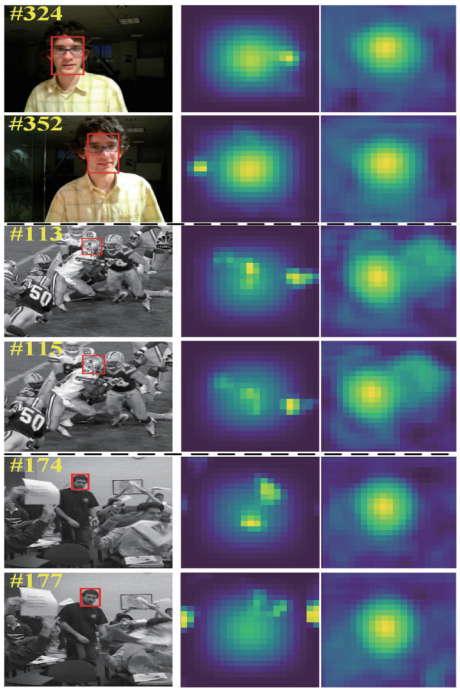
为了验证该模块的有效性, 采用 SiamRPN 作为基线。如表 2 所示, 分别在嵌入 ST 和未嵌入 ST 的情况下来验证 DA 模块。

表 2　DA 模块消融分析

Table 2　Ablation of DA module

跟踪器	ST	DA	OTB100		GOT10k		
			成功率/%	精度	AO	SR_{50}	SR_{75}
Modified SiamRPN			60.70	0.80	0.423	0.426	0.158
		√	**62.40**	**0.81**	**0.458**	**0.543**	**0.251**
	√		63.34	0.83	0.463	0.548	0.254
	√	√	**67.07**	**0.89**	**0.567**	**0.656**	**0.452**

在未嵌入 ST 时嵌入 DA 模块, OTB100 中的成功率提高了 2.8%, GOT10k 中的 AO 提高了 8%。在嵌入 ST 后, 成功率提高了 10%, AO 从 0.463 提高到 0.567。当嵌入 2 个模块时, ST 增强了目标特征, 从而使得 DA 可以得到一个更可靠的判别力掩码, 最终极大提高跟踪器的性能。图 5 中进一步可视化了嵌入 DA 后分类分支输出的响应图。



(a) Frame　(b) Baseline　(c) Baseline+DE

图 5　分类张量响应图

Fig. 5　Classification tensor response map

2.4　对比实验

为了充分衡量所提方案的竞争性。本文在4个测试基准上继续通过对比当前先进的跟踪器来评估DASTSiam跟踪器的优越性。

2.4.1　OTB100对比实验

使用DASTSiam在OTB100测试基准上与其他10个跟踪器进行对比，成功率和精确度曲线如图6和图7所示。从实验结果可以看出，所提模块相对于基线得到极大提升。同时，DASTSiam跟踪器性能超越了DaSiamRPN[16]、SiamDWrpn[13]、GradNet[33]等孪生网络跟踪器。

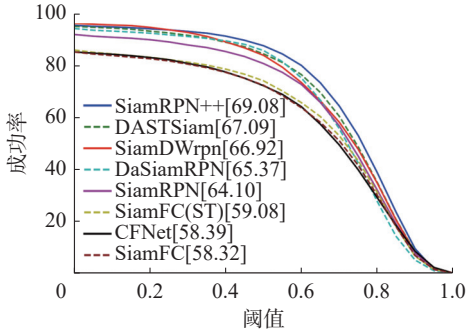


图6　OTB100成功率曲线

Fig. 6　OTB100 success plot

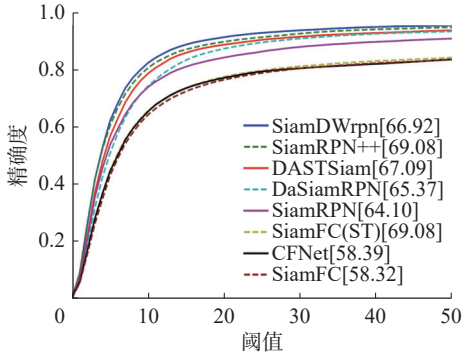


图7　OTB100精确度曲线

Fig. 7　OTB100 precision plot

2.4.2　LaSOT对比实验

使用所提方案在具有挑战的LaSOT测试基准上进行了全面的对比实验。通过图8和图9的成功率曲线和精确度曲线可以看出DASTSiam跟踪器性能优越，超越了SiamMask[34]、SiamRPN++、SPLT[35]、SiamDW等一众先进的孪生网络跟踪算法。

2.4.3　VOT2018对比实验

使用所提方案在VOT2018测试基准上通过不同的评估协议即正确性和鲁棒性来和其他8个先进跟踪算法进行对比实验。如图10所示，DASTSiam相对于基线方案在光照变化、相机移动、移动变化、尺度变化和遮挡几个方面得到极大提升。详细信息参考表3。可以明显看出DASTSiam跟踪器在正确性方面提升较大，因为其可以更加精准地刻画目标的外貌改变。但是鲁棒性方面提升较少，说明所提方案在复杂环境下找回目标的能力有待提高。

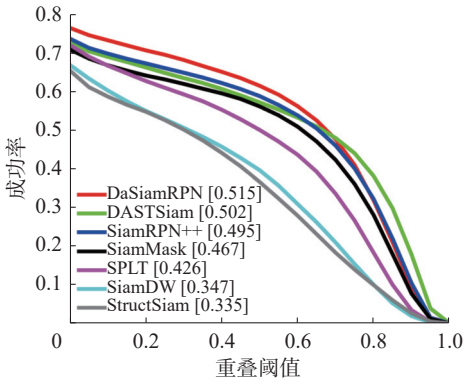


图8　LaSOT成功率曲线

Fig. 8　LaSOT success plot

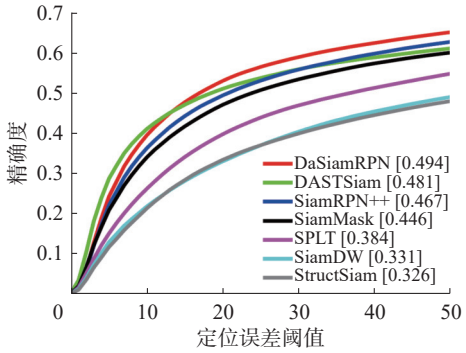


图9　LaSOT精确度曲线

Fig. 9　LaSOT precision plot

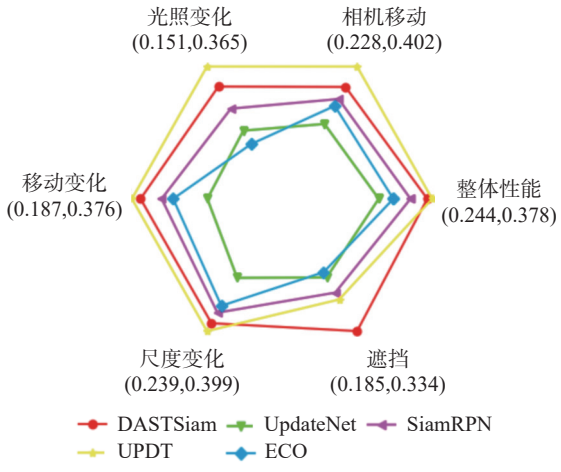


图10　VOT2018对比可视化

Fig. 10　VOT2018 comparison visualization

表3　VOT2018对比分析

Table 3　Comparison results of VOT2018

跟踪器	A	R	EAO
UpdateNet[18]	0.518	0.454	0.244
UPDT[36]	0.536	0.184	0.378
SiamRPN[6]	0.576	0.323	0.324

续表 3

跟踪器	A	R	EAO
ECO[37]	0.484	0.276	0.280
DASTSiam(Ours)	0.585	0.295	0.366

2.4.4 GOT10k 对比实验

使用所提方案在 GOT10k 测试基准与其他先进跟踪器进行对比实验。如表 4 所示, 在 3 个指标 AO、SR_{50}、SR_{75} 下, DASTSiam 跟踪器相对于基线方案有了极大提升, 性能超越了 ECO[37]、SiamRPN++[14]、ATOM[38] 等跟踪器。

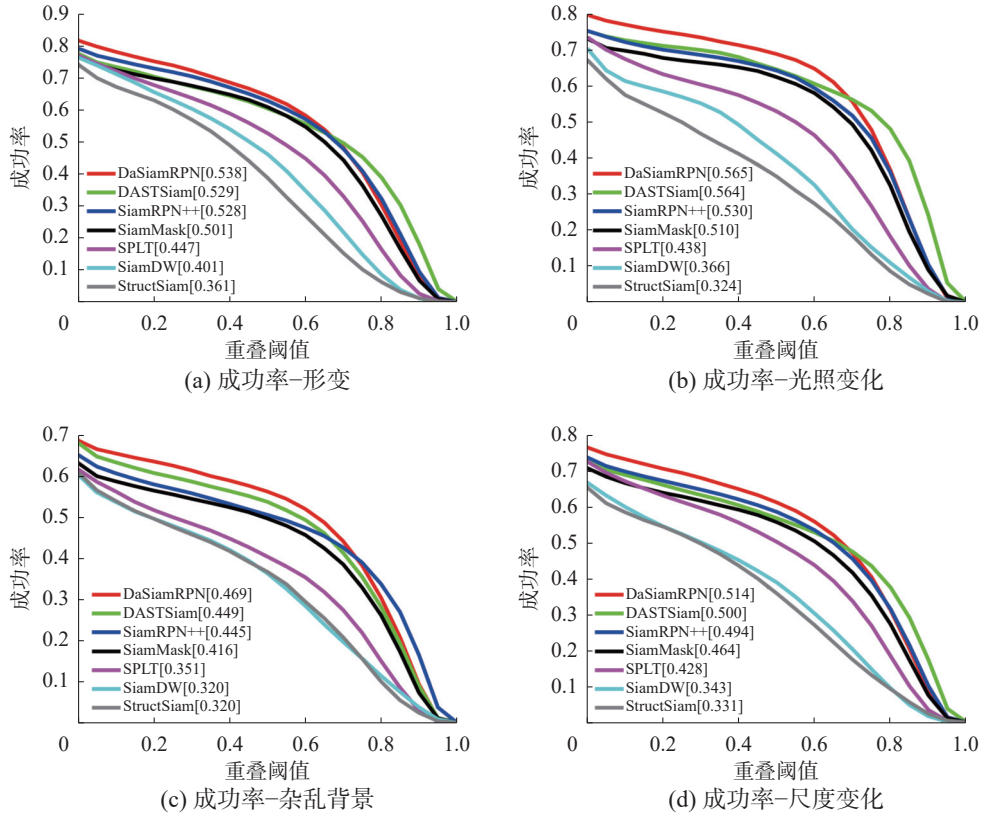
2.5 鲁棒性评估

为了验证 DASTSiam 跟踪解决典型跟踪问题的能力, 在 LaSOT 测试集上执行了全面的鲁棒性评估。如图 11 所示, 与基线相比所提方案在背景杂乱、光照变化、形变、尺度变化等 4 个挑战属性上有极大的提升。验证了 ST 模块在应对目标外貌改变以及 DA 模块在应对杂乱背景下的高鲁棒性。

表 4　GOT10k 对比分析

Table 4　Comparison results of GOT10k

跟踪器	AO	SR_{50}	SR_{75}
ECO[37]	0.316	0.309	0.111
SiamRPN++[14]	0.517	0.615	0.329
ATOM[38]	0.556	0.634	0.402
SiamCAR[15]	0.569	0.670	0.415
DASTSiam(Ours)	0.567	0.656	0.452



(a) 成功率–形变　(b) 成功率–光照变化

(c) 成功率–杂乱背景　(d) 成功率–尺度变化

图 11　LaSOT 鲁棒性分析可视化

Fig. 11　LaSOT robustness analysis

3 结束语

本文提出一个利用时空融合模块和判别力增强模块来应对跟踪的目标外貌变化以及杂乱背景等问题。实验在 4 个基线数据集上充分验证了提出的模块对跟踪性能提升的有效性。总之, 本文通过提出有效的解决方案解决现有孪生网络跟踪器在利用时空信息和从杂乱背景中区分目标 2 方面的局限性。但是目前工作由于网络限制, 缺乏内生旋转不变性, 面对细长物体的旋转变化容易造成跟踪失败。同时目前工作针对腐蚀模板的判断解释性较差, 可能导致跟踪结果漂移。后续将针对这些问题开展进一步研究工作, 以期实现高鲁棒跟踪算法。

参考文献:

[1] 韩瑞泽, 冯伟, 郭青, 等. 视频单目标跟踪研究进展综述[J]. 计算机学报, 2022, 45(9): 1877−1907.
HAN Ruize, FENG Wei, GUO Qing, et al. Single object tracking research: a survey[J]. Chinese journal of computers, 2022, 45(9): 1877−1907.

[2] 王梦亭, 杨文忠, 武雍智. 基于孪生网络的单目标跟踪

算法综述 [J]. 计算机应用, 2023, 43(3): 661−673.
WANG Mengting, YANG Wenzhong, WU Yongzhi. Survey of single target tracking algorithms based on Siamese network[J]. Journal of computer applications, 2023, 43(3): 661−673.

[3] 程旭, 刘丽华, 王莹莹, 等. 基于多帧一致性修正的自监督孪生网络目标跟踪方法 [J]. 计算机学报, 2022, 45(12): 2544−2560.
CHENG Xu, LIU Lihua, WANG Yingying, et al. A multi-frame consistency correction based self-supervised Siamese network method for object tracking[J]. Chinese journal of computers, 2022, 45(12): 2544−2560.

[4] 周春月, 颜巧. 基于高分辨率孪生网络的单目标追踪算法 [J]. 北京交通大学学报, 2020, 44(5): 104−110.
ZHOU Chunyue, YAN Qiao. Single object tracking algorithm based on high-resolution Siamese network[J]. Journal of Beijing Jiaotong University, 2020, 44(5): 104−110.

[5] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional Siamese networks for object tracking[C]//Lecture Notes in Computer Science. Cham: Springer, 2016: 850−865.

[6] LI Bo, YAN Junjie, WU Wei, et al. High performance visual tracking with Siamese region proposal network [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 8971−8980.

[7] GUO Qing, FENG Wei, ZHOU Ce, et al. Learning dynamic Siamese network for visual object tracking[C]// 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017: 1781−1789.

[8] ZHANG Yunhua, WANG Lijun, QI Jinqing, et al. Structured siamese network for real-time visual tracking[C]// European Conference on Computer Vision. Cham: Springer, 2018: 355−370.

[9] 程语嫣, 张九根, 杨圣伟. 多特征融合和尺度适应的相关滤波跟踪算法 [J]. 计算机工程与设计, 2020, 41(12): 3444−3450.
CHENG Yuyan, ZHANG Jiugen, YANG Shengwei. Correlation filtering tracking algorithm based on multi-feature fusion and scale adaptation[J]. Computer engineering and design, 2020, 41(12): 3444−3450.

[10] 茅正冲, 沈雪松. 基于多特征融合的相关滤波跟踪算法 [J]. 计算机与数字工程, 2020, 48(11): 2645−2648, 2782.
MAO Zhengchong, SHEN Xuesong. Correlation filter tracking algorithm based on multi-feature fusion[J]. Computer & digital engineering, 2020, 48(11): 2645−2648, 2782.

[11] 蒲磊, 冯新喜, 侯志强, 等. 基于自适应背景选择和多检测区域的相关滤波算法 [J]. 电子与信息学报, 2020, 42(12): 3061−3067.
PU Lei, FENG Xinxi, HOU Zhiqiang, et al. Correlation filter algorithm based on adaptive context selection and multiple detection areas[J]. Journal of electronics & information technology, 2020, 42(12): 3061−3067.

[12] GIRSHICK R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440−1448.

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

[14] LI Bo, WU Wei, WANG Qiang, et al. SiamRPN: evolution of Siamese visual tracking with very deep networks [C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 4277−4286.

[15] GUO Dongyan, WANG Jun, CUI Ying, et al. SiamCAR: Siamese fully convolutional classification and regression for visual tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 6268−6276.

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

[17] WANG Qiang, TENG Zhu, XING Junliang, et al. Learning attentions: residual attentional Siamese network for high performance online visual tracking[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4854−4863.

[18] ZHANG Lichao, GONZALEZ-GARCIA A, VAN DE WEIJER J, et al. Learning the model update for Siamese trackers[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 4009−4018.

[19] ZHU Zheng, WU Wei, ZOU Wei, et al. End-to-end flow correlation tracking with spatial-temporal attention[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 548−557.

[20] LI Feng, TIAN Cheng, ZUO Wangmeng, et al. Learning spatial-temporal regularized correlation filters for visual tracking[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4904−4913.

[21] LUKEŽIC A, VOJÍR T, ZAJC L C, et al. Discriminative correlation filter with channel and spatial reliability[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 4847−4856.

[22] DANELLJAN M, HÄGER G, KHAN F S, et al. Learn-

ing spatially regularized correlation filters for visual tracking[C]//2015 IEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 4310−4318.

[23] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017: 6000–6010.

[24] ZHANG Shifeng, CHI Cheng, YAO Yongqiang, et al. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9756−9765.

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

[26] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[27] DENG Jia, DONG Wei, SOCHER R, et al. ImageNet: a large-scale hierarchical image database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami: IEEE, 2009: 248−255.

[28] FAN Heng, LIN Liting, YANG Fan, et al. LaSOT: a high-quality benchmark for large-scale single object tracking[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 5369−5378.

[29] RUSSAKOVSKY O, DENG Jia, SU Hao, et al. ImageNet large scale visual recognition challenge[J]. International journal of computer vision, 2015, 115(3): 211−252.

[30] WU Yi, LIM J, YANG M H. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland: IEEE, 2013: 2411−2418.

[31] Kristan M , Leonardis A , Matas J , et al. The sixth visual object tracking VOT2018 challenge results[C]//European Conference on Computer Vision Workshops. Cham: Springer, 2018: 3–53.

[32] HUANG Lianghua, ZHAO Xin, HUANG Kaiqi. 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.

[33] LI Peixia, CHEN Boyu, OUYANG Wanli, et al. GradNet: gradient-guided network for visual object tracking[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 6161−6170.

[34] HU Weiming, WANG Qiang, ZHANG Li, et al. SiamMask: a framework for fast online object tracking and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(3): 3072−3089.

[35] YAN Bin, ZHAO Haojie, WANG Dong, et al. 'skimming-perusal' tracking: a framework for real-time and robust long-term tracking[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 2385−2393.

[36] BHAT G, JOHNANDER J, DANELLJAN M, et al. Unveiling the power of deep tracking[C]//European Conference on Computer Vision. Cham: Springer, 2018: 493−509.

[37] DANELLJAN M, BHAT G, KHAN F S, et al. ECO: efficient convolution operators for tracking[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 6931−6939.

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

作者简介：

黄昱程，硕士研究生，主要研究方向为计算机视觉、目标跟踪。E-mail：1063439128@qq.com。

肖子旺，硕士研究生，主要研究方向为计算机视觉、目标检测。E-mail：107552103759@stu.xju.edu.cn。

艾斯卡尔·艾木都拉，教授，博士生导师，主要研究方向为语音识别与合成、模式识别与图像处理、自然语言处理。登记软件著作权 40 余项，发表学术论文 200 余篇。E-mail：askar@xju.edu.cn。