第 42 卷 第 15 期/2022 年 8 月/光学学报 研究论文

光学学报

基于动态特征注意模型的三分支网络目标跟踪

张子烁[1,2]，宋勇[1,2*]，杨昕[1,2]，赵宇飞[1,2]，周雅[1,2]

[1]北京理工大学光电学院，北京 100081；

[2]精密光电测试仪器及技术北京市重点实验室，北京 100081

摘要 针对实际场景中跟踪目标的快速移动、光照变化和尺度变换等问题，提出一种基于动态特征注意模型(DFA)的三分支网络目标跟踪算法，包括：以SiamRPN++跟踪框架为基础，设计具有动态模板分支的在线更新三分支网络，以强化网络提取特征的语义信息，提高模板特征与搜索目标的匹配相似性；设计面向三分支网络训练的样本生成方法，以改变负样本分配方式，提升正、负样本训练的平衡性；设计一种DFA，通过等效自注意和互注意操作增强模板的历史动态特征，实现模板特征的自适应细化，同时利用通道注意力得分控制搜索特征图的权重分配，提高得分图对目标的响应。相对SiamRPN++、SiamBAN等对比算法，所提算法在包含运动模糊、明暗变化和相似背景干扰等场景的OTB100、VOT2018数据集上，获得了最高成功率(71.0%)和最优鲁棒性(0.122)，同时可满足实时目标跟踪的要求。

关键词 机器视觉；目标跟踪；孪生神经网络；注意力机制

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/AOS202242.1515001

Triplet Network Based on Dynamic Feature Attention for Object Tracking

Zhang Zishuo[1,2], Song Yong[1,2*], Yang Xin[1,2], Zhao Yufei[1,2], Zhou Ya[1,2]

[1]*School of Optics and Photonics, Beijing Institute of Technology, Beijing* 100081, *China*;

[2]*Beijing Key Laboratory for Precision Optoelectronic Measurement Instrument and Technology, Beijing* 100081, *China*

Abstract Considering the fast motion, illumination variation, and scale transform of tracking targets in actual scenarios, a triplet network based on a dynamic feature attention (DFA) model for object tracking is proposed to solve these problems. Specifically, on the basis of the SiamRPN++ tracking framework, an online update triplet network with dynamic template branches is designed to strengthen the semantic information of extracted features and improve the matching similarity between template features and search features. A sample generation method for the triplet network training is developed to change the allocation of negative samples and improve the balance of positive and negative training samples. Moreover, a DFA model, where the historical dynamic features of the templates are enhanced through equivalent self-attention and mutual attention operation, is designed to achieve the adaptive refinement of template features. Meanwhile, the channel attention score is used to control the weight distribution of the search feature maps, and the response of the score maps is improved. Compared with the state-of-the-art algorithms such as SiamRPN++ and SiamBAN, the proposed algorithm has achieved the highest success rate (71.0%) and the best robustness (0.122) on the OTB100 and VOT2018 datasets that contain scenes with motion blur, illumination variation, and similar background interference. This algorithm also can meet the requirement of real-time target tracking.

Key words machine vision; object tracking; Siamese neural network; attention mechanism

1 引 言

目标跟踪在视频监控、人机交互和视觉导航等领域中具有广泛的应用[1-2]。在复杂条件下的目标跟踪中，目标运动、相似背景干扰等因素会导致目标形状、尺度和光照等变化，从而使目标跟踪精度下降。目前，复杂条件下的目标跟踪成为计算机视觉领域中具有挑战性的研究方向。

收稿日期：2022-01-13；**修回日期**：2022-02-27；**录用日期**：2022-03-07

基金项目：国家自然科学基金(81671787)、空间光电测量与感知实验室开放基金课题资助项目(LabSOMP-2018-03)

通信作者：*yongsong@bit.edu.cn

常规目标跟踪主要利用基于相关滤波器的算法，通过提取目标特征训练相关滤波器，进而得到响应得分图。此类算法的速度较快，但由于存在边界效应问题，故其判别器不稳定，精度较低，难以满足复杂条件下的目标跟踪要求[3-4]。同时，基于深度学习的算法利用卷积神经网络（CNN）提取目标不同深度的卷积特征，具有较高的目标辨识能力。然而，此类算法需通过迭代的方式求解，计算复杂度较高，难以满足目标跟踪的实时性要求[5]。

此外，近年来出现的孪生网络目标跟踪算法因同时具有准确度和速度方面的优势而备受研究者关注。Bertinetto等[6]提出了全卷积孪生网络单目标跟踪算法SiamFC算法，此类模板与搜索特征图匹配的跟踪框架打破了基于深度学习的目标跟踪算法无法满足实时性的局限。SiamRPN算法[7]通过网络内的区域候选网络（RPN）[8]自适应微调获得更加精准的边界框。DaSiamRPN算法[9]中增加了干扰感知模型，同时采用局部到全局的重检测策略，可实现长时间且稳定的目标跟踪。UPDT算法[10]令深浅层特征图自适应融合，基于新的质量评估方法得到最佳的目标定位结果。SiamDW算法[11]和SiamRPN＋＋算法[12]均通过加深、加宽孪生网络的主干结构，实现更深层次、更丰富的特征信息提取。UpdateNet算法[13]利用改进的在线更新模块，通过跳跃连接的方式进行残差学习，实现模板自适应更新。SiamAttn算法[14]通过引入变形注意模块提升目标与搜索图像的判别力。SiamBAN算法[15]利用目标边界框自适应回归策略，提升算法对目标边界框的尺寸适应性。

上述孪生网络目标跟踪算法虽然在满足实时跟踪的前提下具有一定的目标跟踪器精度，但是对图像变化特征持续提取的稳定性较低，其在复杂条件下的跟踪精度仍存在较大的提升空间[16]。

基于上述分析，本文提出一种基于动态特征注意模型（DFA）的三分支网络（DFA-TriNet）。本文算法的主要思想如下。

1）设计一种具有动态模板分支的在线更新三分支网络。利用历史目标特征丰富网络提取特征的语义信息，防止跟踪器在预测结果时可能产生的灾难性漂移，提升跟踪器的在线适应能力。

2）设计面向三分支网络训练的样本生成方法，以完善负样本的分配方式，提升正、负样本训练的平衡性。

3）设计一种包括动态模板注意模块和互注意模块的DFA模型，分步融合初始帧模板信息、历史目标信息和当前帧搜索图信息，增强网络提取特征的稳定性，提升跟踪器在复杂条件下的鲁棒性。

最终，提高分类得分对目标位置的响应能力，增强回归得分对目标尺寸的适应性，从而提升跟踪器的精度和鲁棒性，实现在复杂条件下的稳定、实时目标跟踪。

2　所提方法

基于DFA的三分支网络目标跟踪算法的结构如图1所示，主要包括具有动态模板分支的在线更新三分支网络、DFA和目标边界框自适应回归头部网络。其中，DFA包括动态模板注意模块（DFA_update）和互注意模块（DFA_cross）。

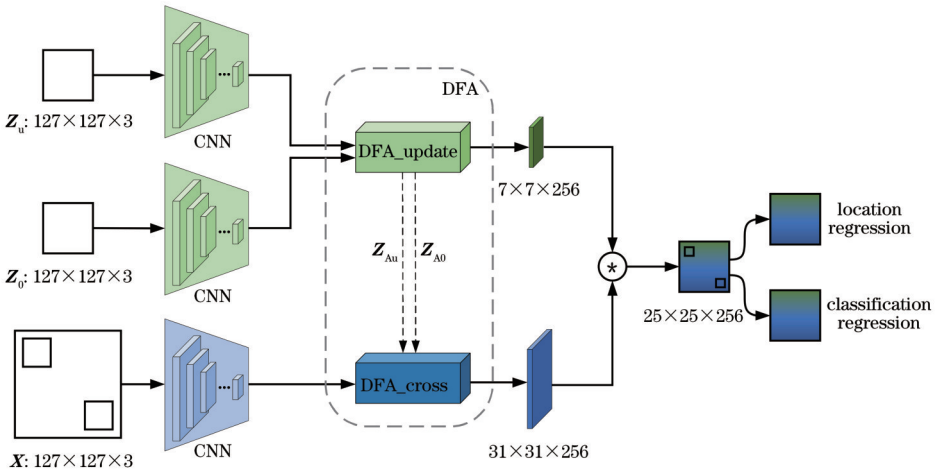

图1　DFA-TriNet结构示意图

Fig. 1　Schematic diagram of DFA-TriNet

2.1　SiamRPN++跟踪框架

孪生网络跟踪算法由共享权重的双分支卷积神经网络组成，能够实现端到端离线训练，以进行网络模型的在线跟踪。孪生网络的两个分支分别对初始帧模板图像和当前帧搜索图像进行特征提取，两幅深度特征图经过互相关操作后可得到相似性响应得分图，对该得分图进行计算可得到目标的位置。常规的全卷积孪生网络相似度函数表示为

$$f(\boldsymbol{X},\boldsymbol{Z})=\varphi_i(\boldsymbol{X})*\varphi_i(\boldsymbol{Z}),\tag{1}$$

式中：$\boldsymbol{X}$为当前帧搜索图像；$\boldsymbol{Z}$为初始帧模板图像；$\varphi_i(\boldsymbol{X})$和$\varphi_i(\boldsymbol{Z})$分别为经过第$i$层后输出的搜索特征图和模板特征图；$*$为互卷积运算；$f(\boldsymbol{X},\boldsymbol{Z})$为相似性响应得分图。

在全卷积孪生网络的基础上，SiamRPN++算法结合RPN模块，将头部网络分为分类分支和回归分支，使网络的定位能力和回归能力分别得到训练。同时，设计一种分层特征聚合结构，融合深浅层网络提取的特征图，以充分利用其上下文语义信息。首先，SiamRPN++的输入模板图 $\boldsymbol{Z}$ 和搜索图 $\boldsymbol{X}$ 在经过ResNet50骨干网络提取特征后可得到深浅层输出特征图。然后，将深浅层输出特征图分别输入RPN模块中，进行权重加和，得到待匹配特征图。最后，通过互相关计算获得响应得分图，经卷积操作后可得到分类得分图和回归得分图。SiamRPN++的相似度函数可以表示为

$$f(\boldsymbol{X},\boldsymbol{Z})=\sum_{i=3}^{5}\varphi_i(\boldsymbol{X})*\sum_{i=3}^{5}\varphi_i(\boldsymbol{Z})。\tag{2}$$

2.2 三分支网络

在常规孪生网络跟踪算法中，模板图像通常固定为对第一帧图像进行裁剪得到的目标区域。由于模板图像中的目标信息（目标模糊、明暗和旋转姿态等特征）在跟踪全程均保持不变，故跟踪器难以适应目标的外观和尺度变换[13]。针对这一问题，本文设计一种具有动态模板分支的三分支网络，可使网络学习目标随时间变化的历史特征信息，同时确保第一帧模板图像内目标稳定特征信息不流失，使动态模板分支具有稳定可靠的在线更新能力。

如图1所示，所提算法的三分支网络具有动态模板分支。在在线跟踪过程中，利用DFA持续地将历史目标信息分别与原始的模板特征图和搜索特征图进行融合。根据第 $t-k$ 帧的跟踪预测结果 $\left(c_x,\ c_y,\ w_{t-k},\ h_{t-k}\right)$，将第 $t-k$ 帧图像裁剪为大小为127×127×3的模板图像作为第 t 帧动态模板分支的输入，其中 $\left(c_x,\ c_y\right)$ 为目标边界框中心点位置坐标，w_{t-k} 和 h_{t-k} 为目标边界框的宽和高。动态模板分支的特征提取网络与原有的模板分支结构和搜索分支结构相同且共享权重。保留第一帧模板图像保证了特征匹配过程的稳定性，融入在线更新的模板图像保证了特征匹配过程的准确性。因此，所提算法在融入历史帧（第 $t-k$ 帧）图像目标特征的同时，仍保留稳定的初始帧图像目标特征，从而使跟踪器获得了稳健的在线学习目标变化的能力。

针对该三分支网络，本文设计相应的离线训练样本分配方式。在训练过程中，网络一次输入的样本组包括尺寸为127×127×3的初始帧模板图像、尺寸为127×127×3的动态模板图像和尺寸为255×255×3的搜索图像。为了提升网络对目标的辨别能力，防止跟踪过程中相似物、遮挡等干扰，引入负样本对，即至少一种模板图像与搜索图像来自不同的视频序列。

正负样本分配方式如表1所示。若初始帧模板图与搜索图来自同一视频序列，则两张图像距离不能超过100帧。若动态模板图与搜索图来自同一视频序列，则两张图像距离不能超过5帧。若任意模板图与搜索图来自不同视频序列，则在搜索图所在序列外的其他序列中，随机抽取一张图像作为模板图。其中，负样本有三种模板图与搜索图不匹配的分配方式，三者在总负样本中的占比相同。正负样本组数据量的比例为8∶2[9]。

表1 DFA-TriNet视频序列训练数据正负样本分配方式

Table 1 Positive and negative sample allocation of training data in video sequence with DFA-TriNet

Sample type	Initial frame template image (within 100 frames)	Update template image (within 5 frames)
Positive sample	Same	Same
Negative sample (equally distributed)	Same	Different
	Different	Same
	Different	Different

2.3 动态特征注意模型

虽然SiamRPN++结合了卷积神经网络深浅层提取到的不同层面的特征信息，但是特征图包含的语义信息仍不能很好地判别目标和相似背景，进而导致其在复杂条件下跟踪结果的漂移[13]。为了提升网络对目标的判别能力，本文设计了DFA，其包含动态模板注意模块和互注意模块。两个模块均基于在线更新的三分支网络结构设计，可充分融合历史目标特征，实现跟踪器对目标随时间变化的动态特征的注意。

2.3.1 动态模板注意模块

为了提升网络对预测框漂移的抵抗力，需保留初始帧模板图像，以提供准确的模板特征[17]。在此基础上，为了进一步丰富模板特征，需融合动态模板特征。由于动态模板与初始帧模板间存在差异但又同时包含目标特征，简单的加法操作难以稳定地融合目标特征信息，故本文设计了动态模板注意模块，其结构如图2所示。该模块包括对两幅模板特征图的自注意操作和互注意操作，关注动态模板特征对目标的权重再分配，实现模板特征的在线自适应更新。

首先，利用多层感知混合器（Mixer_Block）对初始帧模板特征和动态模板特征的空间位置和特征通道进行全连接操作处理，利用自身信息加强模板特征图。经过多层感知混合器后的模板特征 $\boldsymbol{Z}_{\mathrm{Au}}$ 和 $\boldsymbol{Z}_{\mathrm{A0}}$，通过非线性激活函数Sigmoid分别与原模板特征 $\boldsymbol{Z}_{\mathrm{u},i}$ 和 $\boldsymbol{Z}_{0,i}$ 进行元素级乘法获得 $\boldsymbol{Z}_{\text{self-u}}$ 和 $\boldsymbol{Z}_{\text{self-0}}$，使用两次在不同维度上的多层感知操作，实现模板特征的空间自注意和通道自注意，以突出有效信息，压缩无效信息，提升网络对目标特征的判别能力。

多层感知混合器的结构如图3所示，利用多层感知器（MLP）进行空间混合和通道混合。空间混合操

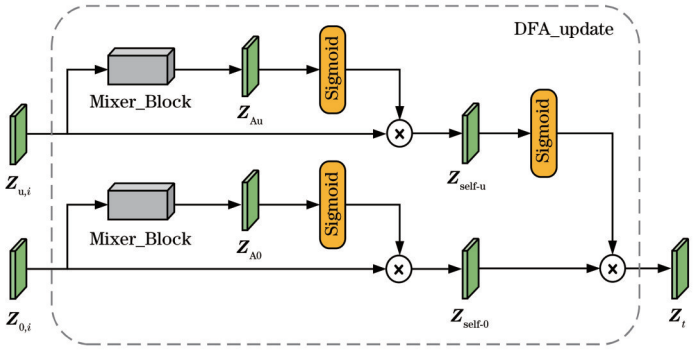

图 2　DFA_update 模块结构图

Fig. 2　Structural diagram of DFA_update module

作可融合模板特征各通道的展平空间信息，通道混合操作可融合模板特征各空间分布位置的通道信息，并利用残差法跳跃连接空间混合与通道混合的结果。多层感知混合器针对分类任务的可靠性已被实验验证[18]，利用空间分布的重塑操作（Reshape）代替嵌入层（Embedding），能够在使用更少计算量的前提下实现与 ViT 算法[19]性能相当的特征自注意效果[18]。

多层感知混合器的计算操作可描述为

$$\begin{cases}\overline{Y}=\overline{X}+L_2\delta\left[L_1M_{\text{LayerNorm}}\left(\overline{X}\right)\right]\\ \overline{Z}=\overline{Y}+L_4\delta\left[L_3M_{\text{LayerNorm}}\left(\overline{Y}\right)\right]\end{cases},\tag{3}$$

式中：$\overline{X}$、$\overline{Y}$ 和 $\overline{Z}$ 为展平的模板特征图；$L_n(n=1,2,3,4)$ 为不同维度的线性变换；$\delta(\cdot)$ 为激活函数 GELU；$M_{\text{LayerNorm}}(\cdot)$ 为层归一化。

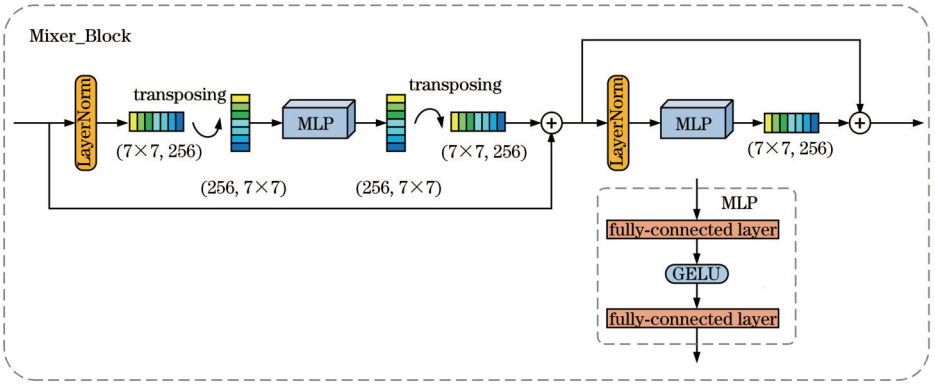

图 3　多层感知混合器结构图

Fig. 3　Structural diagram of multilayer perceptual mixer block

然后，利用非线性激活函数 Sigmoid 对动态模板特征进行处理，获得基于历史变化模板特征的再分配权重，再分配权重与初始帧模板特征 $Z_{\text{self-0}}$ 经元素级乘法融合，利用互注意操作融合两幅特征图，得到用于特征匹配的模板特征图 Z_t，以实现模板特征的自适应细化。

2.3.2　互注意模块

在常规孪生网络跟踪算法中，目标特征与搜索区域特征计算过程相互独立，且仅通过一次互相关操作来进行特征匹配，这限制了网络对目标特征的学习能力[20]。针对上述问题，本文设计如图 4 所示的互注意模块，实现模板特征与搜索特征的加权融合，以突出搜索特征图上的目标信息，从而有效聚合与关联模板和搜索区域间的信息。

利用动态模板注意模块计算过程中生成的特征图 Z_{Au} 和 Z_{A0}，依次进行层归一化、平均池化（Avg_Pooling）和重塑操作，得到尺寸压缩为 1×1×256 的动态模板和初始帧模板的特征图，该特征图包含原模板特征图的全局感受野[21]。计算过程为

$$p(Z)=\frac{1}{H\times W}\sum_{m=1}^{H}\sum_{j=1}^{W}Z_{m,j},\tag{4}$$

$$\boldsymbol{\Phi}(Z)=p\left[M_{\text{LayerNorm}}(Z)\right],\tag{5}$$

式中：$p(\cdot)$ 为平均池化操作；Z 为特征图；H 和 W 为特征图的高度和宽度。

尺寸为 1×1×256 的模板特征图 $\boldsymbol{\Phi}(Z)$ 包含各通道维度下的全局空间特征信息，利用非线性激活函数 Sigmoid 对该动态模板特征图进行处理，获得针对通道维度的再分配权重，将该再分配权重与该初始帧模板特征图进行元素级乘法融合。最后，将元素级乘法融合得到的特征图与搜索特征图 $\boldsymbol{X}_t$ 融合，获得具有模板特征的特征图在通道维度上的注意力再分配，即具有通道注意力[22]的搜索特征图 $\boldsymbol{X}_t$。上述过程可表示为

$$\boldsymbol{X}_t=M_{\text{scale}}\left\{\boldsymbol{X}_t,M_{\text{scale}}\left[\boldsymbol{\Phi}\left(Z_{\text{A0}}\right),\sigma\boldsymbol{\Phi}\left(Z_{\text{Au}}\right)\right]\right\},\tag{6}$$

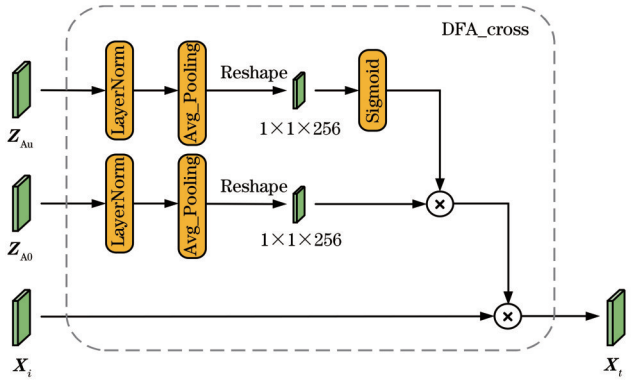

图 4 DFA_cross模块的结构图

Fig. 4 Structural diagram of DFA_cross module

式中：$M_{\text{scale}}(\cdot)$为元素级乘法融合，即对特征通道进行重新加权；σ为非线性激活函数Sigmoid。

利用DFA获得的模板特征图$\boldsymbol{Z}_t$和搜索特征图$\boldsymbol{X}_t$，结合式(2)可得到跟踪网络的响应得分图，该响应得分图可用于后续的分类预测和回归预测。

2.4 目标边界框自适应回归

由于RPN模块的存在，故SiamRPN++算法需要预设多个固定尺寸的锚框，用于网络预测候选框的辅助回归。在目标跟踪过程中，当网络预测的候选框数量较大时，将导致RPN模块的计算量庞大，同时超参数的设置会导致网络对先验知识过于敏感。为保证所提算法的跟踪实时性和精确性，提出用无锚框的目标边界框自适应回归策略[15]代替传统的RPN模块。

在所提算法的头部网络中，采用无锚框的目标边界框自适应回归策略对结果进行预测，无需预设含有超参数的锚框，在减少计算量的同时，减少网络对边界框尺寸信息先验知识的敏感性，提升网络预测框的自由度。将分类得分图中得分最高的位置视为目标位置，将回归得分图对应目标位置的得分$(l,\ t,\ r,\ b)$视为目标边界框的尺寸偏移量。目标边界框在响应图上的位置$(\tilde{x}_1,\ \tilde{y}_1,\ \tilde{x}_2,\ \tilde{y}_2)$由目标边界框的尺寸偏移量计算得到，具体计算公式为

$$\begin{cases}\tilde{x}_1 = p_m - l\\ \tilde{y}_1 = p_j - t\\ \tilde{x}_2 = p_m + r\\ \tilde{y}_2 = p_j + b\end{cases},\tag{7}$$

式中：$(p_m,\ p_j)$为响应图上目标的位置坐标。

在训练过程中，分类得分由交叉熵损失函数进行计算，回归得分由CIoU损失函数[23]进行计算，CIoU损失函数相较IoU损失函数[24]增加了目标框距离和长宽比的惩罚项，充分体现了预测框与目标框之间的重叠面积、中心点距离和长宽比，具体公式为

$$L_{\text{CIoU}} = 1 - L_{\text{IoU}} + \frac{\rho^2(b, b_{\text{gt}})}{c^2} + \alpha v,\tag{8}$$

式中：L_{CIoU}和L_{IoU}分别为CIoU损失函数和IoU损失函数；$\rho^2(b, b_{\text{gt}})$为预测框中心点b与真实框中心点b_{gt}间的欧氏距离。c为包含预测框面积与真实框面积的最小长方形对角线距离。α与v的计算公式为

$$\alpha = \frac{v}{1 - L_{\text{IoU}} + v},\tag{9}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{g_{\text{w}}}{g_{\text{h}}} - \arctan\frac{w}{h}\right)^2,\tag{10}$$

式中：g_{w}和g_{h}为真实框的宽和高；w和h为预测框的宽和高。

最终，总损失由分类损失与回归损失按等比例加和计算。

3 实验分析

3.1 实验设置

所提算法的实验硬件环境为多核处理器为Intel Xeon Silver 4214 CPU，内存为128 G，主频为2.20 GHz，显卡为NVIDIA GeForce RTX 2080ti GPU。编程环境为Pytorch深度学习框架的python3.7。

3.1.1 训练参数设置

所提算法在上述平台上进行训练，使用目标跟踪数据集ImageNet VID[25]、LaSOT[26]、YouTube Bounding Boxes[27]和用于补充正样本的目标检测数据集MS COCO[28]、ImageNet DET[25]。在训练过程中，使用在ImageNet数据集上预训练后的权重对算法的骨干网络进行初始化。利用随机梯度下降优化器(SGD)对网络进行训练，权重衰减率设置为0.0001，动量设置为0.9，小批次训练样本数量设置为14组，每轮训练样本数量为6×10^5组，训练迭代轮数epoch为40。训练的初始学习率定为1×10^{-3}，训练前10轮为热身训练阶段，设置学习率从1×10^{-3}升至5×10^{-3}，后30轮设置学习率从5×10^{-3}降至5×10^{-5}，且以指数衰减。在前20轮中，固定骨干网络权重，仅训练动态特征注意模块和目标边界框自适应回归头部网络。在后20轮中，以当前轮次学习率的1/10对骨干网络进行微调。

3.1.2 更新参数设置

为了有效融合动态特征，需要对动态模板分支

的输入图像进行实时更新。其中，更新帧数频率会影响网络对特征学习的平衡性[29]，更新频率过慢会导致网络学习目标特征变化性不足，更新频率过快会导致网络误学习遮挡物等干扰的特征，从而污染目标特征。

为了确定合适的更新帧数频率，使用VOT2018数据集对每k帧在线更新动态模板图像进行对比实验。在训练过程中，动态模板图像选择搜索图像前后5帧内的随机图像，设置k为1～10的更新帧频，实验结果如表2所示。

表2 基于VOT2018数据集的更新参数实验结果

Table 2 Experimental results for updating parameter based on VOT2018 dataset

k	1	2	3	4	5	6	7	8	9	10
EAO	0.452	0.466	0.463	0.462	0.469	0.468	0.465	0.464	0.465	0.462

根据表2结果，当k为5帧更新频率时，所提网络在测试数据集上的预期平均重叠期望(EAO)最高，可实现最优性能，故设置k=5。

3.2 消融实验

为了验证在线更新三分支网络结构和动态特征注意模块对算法产生的影响，设计如表3所示的消融实验。其中，BA(Box Adaptive)为目标边界框自适应回归策略，UB(Update Branch)为动态模板分支。

对比基线算法：仅替换目标边界框自适应回归策略，EAO结果提升3.4个百分点；添加动态模板分支，但不进行自注意和互注意操作，更新后的模板与原模板直接相加，结果提升4.5个百分点；再利用DFA处理特征图，结果提升5.5个百分点。在跟踪速度上：仅替换目标边界框自适应回归策略，算法时效性得到提升；添加动态模板分支并利用DFA处理特征图，算法时效性虽略有降低，但仍优于基线算法，且满足实时跟踪要求(25 frame/s)。实验结果表明，本文设计的三分支网络和DFA均对算法有较大贡献，整体提升了网络对目标的判别能力。

3.3 定量分析

为了验证所提算法的有效性，使用公共测试数据集OTB100和VOT2018，将所提算法与近几年跟踪性能优异的SiamFC[6]、CFNet[30]、SiamDW[11]、DeepSRDCF[31]、DaSiamRPN[9]和SiamRPN++[12]6种跟踪算法进行对比。

OTB100数据集以一次性评估方式(OPE)下的跟踪成功率和跟踪精度作为跟踪算法的评价指标，具体实验结果如图5所示，图例中的数字为最终的计算结果。

表3 基于VOT2018数据集的消融实验结果

Table 3 Results of ablation experiment based on VOT2018 dataset

Index	SiamRPN++	SiamRPN+++ BA	SiamRPN+++ BA+UB	SiamRPN+++ BA+UB+DFA
EAO	0.414	0.448	0.459	0.469
Tracking speed /(frame·s^{-1})	35	44	38	37

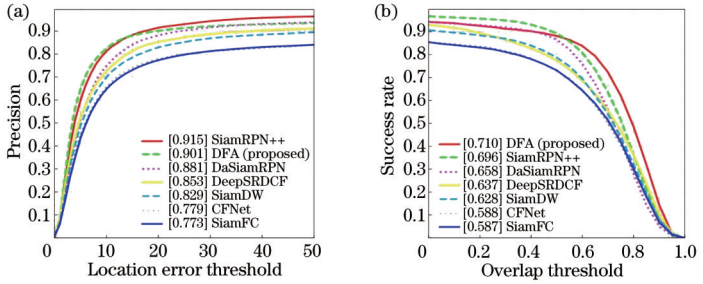

图5 基于OTB100数据集的实验结果。(a)精确度；(b)成功率

Fig. 5 Experimental results based on OTB100 dataset. (a) Precision; (b) success rate

可以看出，所提算法与SiamRPN++算法相比，成功率提升了1.4个百分点，精度降低了1.4个百分点，在OTB全数据集上表现与SiamRPN++相当。

为了更好地测试本文算法的性能，利用OTB100数据集中不同属性类型的图像序列组分别对所提算法和对比算法进行评价，结果如图6所示。

由图6总结得到，在背景杂乱、快速移动、光照变化和运动模糊属性的序列中，所提算法在成功率和精度上均优于SiamRPN++算法，分别最高可提升3.8个百分点和5.0个百分点，表现出了更为优异的跟踪性能。结果表明，对网络输入历史特征信息并实时更新模板注意能够有效提高网络对目标变化的适应性。

在平面内旋转、平面外旋转、出视野和尺度变换属性的序列中，所提算法在成功率上优于SiamRPN++算法，表明所提算法适应性强，在跟踪效果上有一定的提升。

当出现形变、低分辨率和遮挡情况时，所提算法的表现略逊于SiamRPN++算法，但仍能达到较优秀的

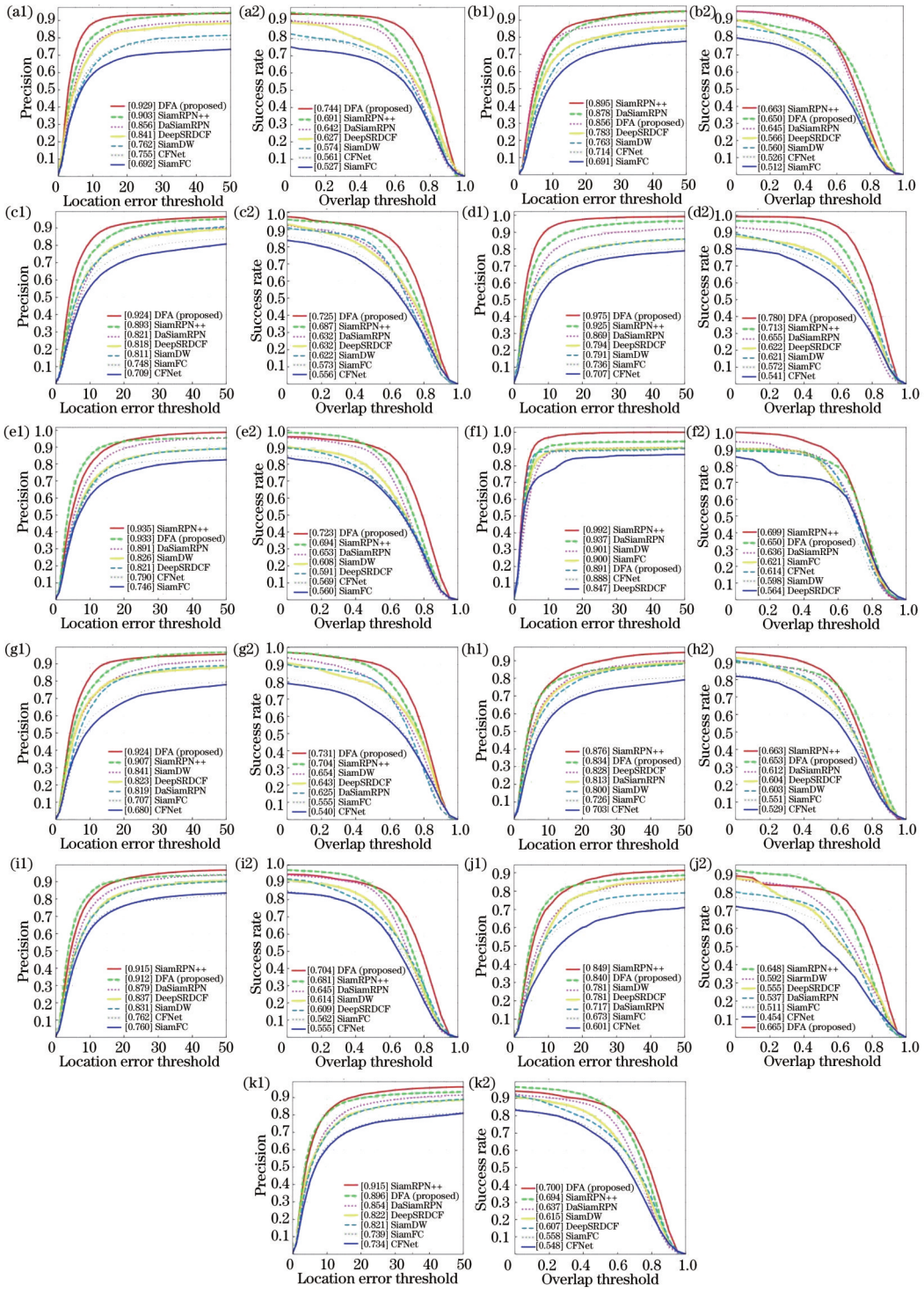

图 6 基于OTB100数据集不同属性序列的实验结果。(a1)(a2)背景杂乱;(b1)(b2)形变;(c1)(c2)快速移动;(d1)(d2)光照变化;(e1)(e2)平面内旋转;(f1)(f2)低分辨率;(g1)(g2)运动模糊;(h1)(h2)遮挡;(i1)(i2)平面外旋转;(j1)(j2)出视野;(k1)(k2)尺度变换

Fig. 6 Experimental results of different attribute sequences based on OTB100 dataset. (a1) (a2) Background clutters; (b1)(b2) deformation; (c1) (c2) fast motion; (d1) (d2) illumination variation; (e1)(e2) in-plane rotation; (f1)(f2) low resolution; (g1)(g2) motion blur; (h1) (h2) occlusion; (i1) (i2) out-of-plane rotation; (j1)(j2) out-of-view; (k1)(k2) scale variation

水平,说明当出现目标特征不明显(低分辨率)或者目标特征剧烈变化(形变和遮挡)时,所提算法对目标特征的学习仍存在提升空间。

VOT2018数据集以一次性评估方式下的跟踪精度、鲁棒性和EAO作为跟踪算法的评价指标,将所提算法与对比算法进行实验结果对比,具体实验结果如表4所示。其中,带有下划横线的数值为最优结果,带有下划波浪线的数值为次优结果。

表 4 基于 VOT2018 数据集的实验结果

Table 4 Experimental results based on VOT2018 dataset

Index	DaSiamRPN	UPDT	SiamRPN	ATOM	UpdateNet	SiamRPN++	SiamBAN	SiamAttn	Proposed
EAO	0.326	0.378	0.383	0.401	0.403	0.414	0.452	0.470	0.469
Accuracy	0.569	0.536	0.586	0.590	0.583	0.600	0.597	0.630	0.614
Robustness	0.337	0.184	0.276	0.203	0.225	0.234	0.178	0.160	0.122
Tracking speed / (frame·s^{-1})	59	0.4	38	30	55	35	44	33	37

由表 4 基于 VOT2018 数据集的实验结果可以看出，对比该数据集上其他最新跟踪算法，所提算法具有优秀的跟踪性能。所提算法以次优准确性(0.614)和最优鲁棒性(0.122)，获得了仅次于 SiamAttn 的 EAO(0.469)。所提算法的在线自适应更新有效地提升了目标跟踪任务的稳定性，动态特征注意模块实现了网络对目标动态历史特征的持续关注，进而提升了算法的整体性能。同时，所提算法速度能够达到 37 frame/s，满足实时跟踪要求。

3.4 定性分析

为了更清晰地验证所提算法性能，对部分图像序列进行可视化结果展示。从 OTB100 中选取 5 个具有各种跟踪难点的视频序列，将所提算法与 SiamFC[6]、CFNet[30]、SiamDW[11]、DeepSRDCF[31]、DaSiamRPN[9]、SiamRPN++[12]6 种跟踪算法进行对比，如图 7 所示。

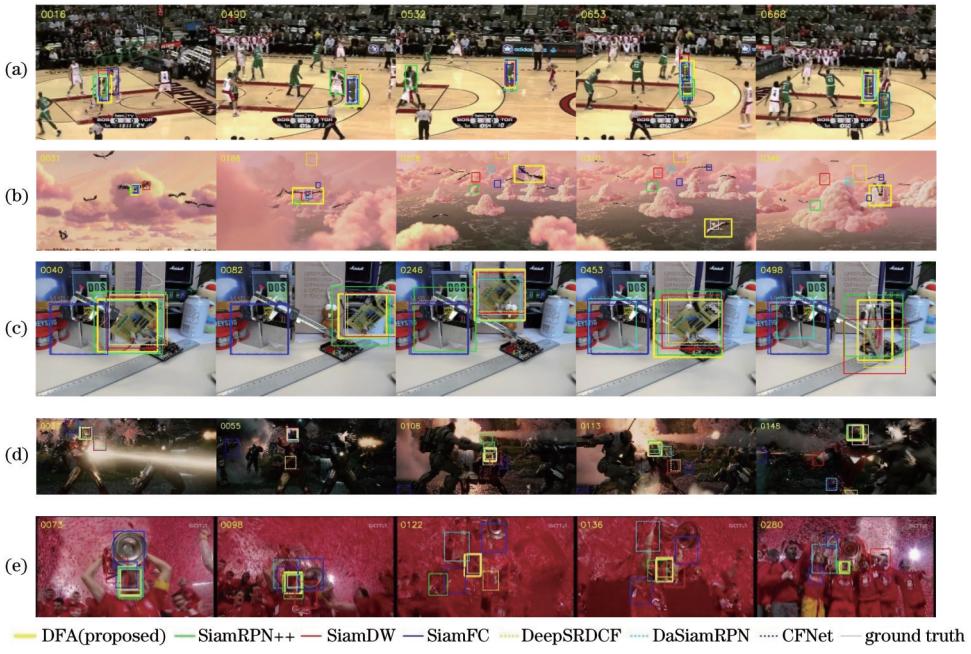

图 7 所提算法和对比算法在不同属性视频序列中的实际跟踪结果图。(a) Basketball；(b) Bird1；(c) Board；(d) Ironman；(e) Soccer

Fig. 7 Actual tracking results of proposed algorithm and comparison algorithms in video sequences with different attributes. (a) Basketball; (b) Bird1; (c) Board; (d) Ironman; (e) Soccer

针对快速移动和明暗变化等情况，可以看出，所提算法较对比算法表现更佳。例如：在 Ironman 序列中，仅有 SiamRPN++与所提算法能够全程成功跟踪，而其他跟踪器会因光照变换而使预测框偏移；在 Soccer 序列中，目标物体快速移动，且因遮挡产生了明暗变化，DFA 能够加强算法对目标特征变换的学习，进而实现全程持续、稳定的跟踪。

针对环境内相似物体的干扰因素，在 Basketball 和 Board 序列中，SiamFC、SiamRPN++等对比算法会被周围的相似物干扰，预测框完全偏移且在一段时间内无法巡回。同时，所提算法的动态特征注意模块能够有效地区分目标与相似物，实现相似物干扰条件下的稳定目标跟踪。

针对目标消失重现问题，由 Bird1 序列可以看出，对比算法在重现后会误判目标的位置信息，而所提算法能够实现持续、稳定的跟踪。在复杂背景、目标剧烈变换或消失重现的场景中，所提算法能够稳定地完成实时跟踪，相对对比算法具有更高的跟踪成功率。

4 结 论

针对目标跟踪任务中物体快速移动、光照变化和尺度变换等复杂条件，提出一种基于 DFA 的三分支网络目标跟踪算法。所提算法以端到端离线训练的 SiamRPN++跟踪框架为基础，设计用于在线模板更

新的三分支网络和相应的训练方式，使网络具有在线更新能力，提升网络提取历史特征信息的动态完整性，以提高模板特征与目标的匹配相似度。设计DFA提升模板特征和搜索特征的判别能力，以实现自适应特征细化，提升跟踪器在复杂条件下的跟踪稳定性。在公共数据集上的对比实验结果表明，所提算法能够更好地适应明暗变化、运动模糊和相似背景干扰等情况：在OTB100上可获得的最高成功率(71.0%)和第二精确率(90.1%)；在VOT2018上可获得最优鲁棒性(0.122)、第二精度(0.614)和第二EAO(0.469)，同时可满足实时目标跟踪要求。综合丢失情况、预测框尺寸适应性和跟踪速率等指标可以发现，所提算法可实现持续、稳定的跟踪，表明基于三分支网络的模板更新和DFA能够有效地提取特征信息。

参 考 文 献

[1] 仇祝令, 查宇飞, 朱鹏, 等. 基于孪生神经网络在线判别特征的视觉跟踪算法[J]. 光学学报, 2019, 39(9): 0915003.
Qiu Z L, Zha Y F, Zhu P, et al. Visual tracking algorithm based on online feature discrimination with Siamese network[J]. Acta Optica Sinica, 2019, 39(9): 0915003.

[2] 李勇, 杨德东, 韩亚君, 等. 融合扰动感知模型的孪生神经网络目标跟踪[J]. 光学学报, 2020, 40(4): 0415002.
Li Y, Yang D D, Han Y J, et al. Siamese neural network object tracking with distractor-aware model[J]. Acta Optica Sinica, 2020, 40(4): 0415002.

[3] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]// 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.

[4] 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.

[5] 孟琭, 李诚新. 近年目标跟踪算法短评：相关滤波与深度学习[J]. 中国图象图形学报, 2019, 24(7): 1011-1016.
Meng L, Li C X. Brief review of object tracking algorithms in recent years: correlated filtering and deep learning[J]. Journal of Image and Graphics, 2019, 24(7): 1011-1016.

[6] 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.

[7] Li B, Yan J J, Wu W, et al. High performance visual tracking with Siamese region proposal network[C]// 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.

[8] 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.

[9] Zhu Z, Wang Q, Li B, et al. Distractor-aware Siamese networks for visual object tracking[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer Vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11213: 103-119.

[10] Bhat G, Johnander J, Danelljan M, et al. Unveiling the power of deep tracking[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11206: 493-509.

[11] Zhang Z P, Peng H W. Deeper and wider Siamese networks for real-time visual tracking[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4586-4595.

[12] Li B, Wu W, Wang Q, et al. SiamRPN++: evolution of Siamese visual tracking with very deep networks[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4277-4286.

[13] Zhang L C, Gonzalez-Garcia A, de Weijer J V, et al. Learning the model update for Siamese trackers[C]// IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 4009-4018.

[14] Yu Y C, Xiong Y L, Huang W L, et al. Deformable Siamese attention networks for visual object tracking [C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6727-6736.

[15] Chen Z D, Zhong B N, Li G R, et al. Siamese box adaptive network for visual tracking[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6667-6676.

[16] Chen X L, He K M. Exploring simple Siamese representation learning[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 15745-15753.

[17] 董吉富, 刘畅, 曹方伟, 等. 基于注意力机制的在线自适应孪生网络跟踪算法[J]. 激光与光电子学进展, 2020, 57(2): 021510.
Dong J F, Liu C, Cao F W, et al. Online adaptive Siamese network tracking algorithm based on attention mechanism[J]. Laser & Optoelectronics Progress, 2020, 57(2): 021510.

[18] Tolstikhin I, Houlsby N, Kolesnikov A, et al. MLP-mixer: an all-MLP architecture for vision[EB/OL]. (2021-05-04)[2021-06-09]. https://arxiv.org/abs/2105.01601.

[19] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition

at scale[C]//International Conference on Learning Representations (ICLR) Oral, May 4, 2021, Vienna, Austria. [S.l.: s.n.], 2021.

[20] Danelljan M, Bhat G, Khan F S, et al. ATOM: accurate tracking by overlap maximization[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 4655-4664.

[21] 李畅, 杨德东, 宋鹏, 等. 基于全局感知孪生网络的红外目标跟踪[J]. 光学学报, 2021, 41(6): 0615002.
Li C, Yang D D, Song P, et al. Global-aware Siamese network for thermal infrared object tracking[J]. Acta Optica Sinica, 2021, 41(6): 0615002.

[22] Hu J, Shen L, Albanie S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.

[23] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: faster and better learning for bounding box regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12993-13000.

[24] Yu J H, Jiang Y N, Wang Z Y, et al. UnitBox: an advanced object detection network[C]//MM '16: Proceedings of the 24th ACM international conference on Multimedia, October 15-19, 2016, Amsterdam, The Netherlands. New York: ACM Press, 2016: 516-520.

[25] 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.

[26] Fan H, Lin L T, Yang F, et al. LaSOT: a high-quality benchmark for large-scale single object tracking[C]//IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 5369-5378.

[27] Real E, Shlens J, Mazzocchi S, et al. YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 7464-7473.

[28] Lin T Y, Maire M, Belongie S J, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[29] Zhao F, Zhang T, Song Y B, et al. Siamese regression tracking with reinforced template updating[J]. IEEE Transactions on Image Processing Society, 2021, 30: 628-640.

[30] Valmadre J, Bertinetto L, Henriques J, et al. End-to-end representation learning for correlation filter based tracking[C]//IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 5000-5008.

[31] Danelljan M, Häger G, Khan F S, et al. Convolutional features for correlation filter based visual tracking[C]//IEEE International Conference on Computer Vision Workshop, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 621-629.