第 58 卷 第 20 期/2021 年 10 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

结合缓冲区与三元组损失的孪生网络目标跟踪

郭嘉，王鹏*，杨永侠，李晓艳，邸若海，李雪

西安工业大学电子信息工程学院，陕西 西安 710021

摘要 针对 SiamRPN(Siamese Region Proposal Network)在目标被短时遮挡以及外观剧烈变化的情况下存在定位不准确的问题，提出一种结合目标跟踪缓冲区与三元组损失的目标跟踪算法。该算法首先将原有的固定模板改为动态模板，提升复杂环境下相似度判别的可靠性；然后在模板缓冲区稀疏地缓存目标外观以应对跟踪过程中非语义样本的干扰，增强目标跟踪的鲁棒性；最后应用三元组损失以充分利用目标的正负样本特征，使跟踪更加具有判别能力。使用 OTB100 数据集进行实验，结果表明所提算法的成功率曲线下面积较 SiamRPN 提高了 0.021，平均中心位置误差降低了 25.56 pixel，平均重叠率提高了 25.2%。

关键词 机器视觉；孪生网络；区域提议网络；缓冲区；三元组损失

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202158.2015002

Siamese Network Target Tracking Based on Buffer and Triplet Loss

Guo Jia, Wang Peng*, Yang Yongxia, Li Xiaoyan, Di Ruohai, Li Xue

School of Electronic and Information Engineering, Xi'an Technological University, Xi'an, Shaanxi 710021, *China*

Abstract Aiming at the problem of inaccurate positioning of the SiamRPN (Siamese Region Proposal Network) when the target is temporarily blocked and the appearance changes drastically, a target tracking algorithm combining target tracking buffer and triple loss is proposed. First, the original fixed template is changed into dynamic template to improve the reliability of similarity discrimination in complex environment. Then, the image of the target is sparsely cached in the template buffer to deal with the interference of non-semantic samples in the process of tracking and enhance the robustness of target tracking. Finally, the triplet loss is applied to make full use of the positive and negative sample characteristics of the target to make the tracking more discriminant. Experimental results with OTB100 dataset show that compared with SiamRPN, the area under the success curve of the improved algorithm increases by 0.021, the average center position error decreases by 25.56 pixel, and the average overlap rate increases by 25.2%.

Key words machine vision; Siamese network; region proposal network; buffer module; triplet loss

OCIS codes 110.2960; 150.1135; 110.2970

1 引 言

目标跟踪是计算机视觉领域中的热点问题之一，目标跟踪任务通常根据给定的第一帧图像中的目标来预测后续视频序列中目标出现的位置，该技术在人机交互和视频监控等领域得到广泛的应用，但是跟踪过程中也面临着许多挑战，如目标遮挡、尺度变化、旋转和形态变化都会对目标的跟踪能力产生影响。

近年来，基于孪生网络的算法因其具有精度高与实时性强的特点而逐渐成为目标跟踪研究领域的热门研究方向。2016 年，Bertinetto 等[1]提出了基

收稿日期：2020-12-08；**修回日期**：2020-12-26；**录用日期**：2021-01-07

基金项目：国家自然科学基金(61671362)、陕西省科技厅重点研发计划(2019GY-022)、西安市科技计划(2020KJRC0037)、西安市未央区科技计划(201923)、西安工业大学校长基金面上培育项目(XGPY200217)

通信作者：* wang_peng@xatu.edu.cn

于离线端到端训练的全卷积孪生网络的跟踪方法，即 SiamFC(Fully-Convolutional Siamese Networks)，其跟踪速度在 GPU(Graphic Processing Units)上可以达到 86 frame/s，同时性能上也超过绝大多数的实时跟踪器。随后，针对这一方法不断进行改进，Valmadre 等[2]提出了 CF-Net(Correlation Filter Networks)，其是在 SiamFC 的基础上加入可以端到端训练的相关滤波器层，在轻量级架构中可以实现高精度的运行。2018 年，Dong 等[3]在 SiamFC 的基础上加入了三元组损失函数，有效提升了跟踪过程中目标的判别能力。为了应对目标定位问题，Li 等[4]提出了 SiamRPN(Siamese Region Proposal Network)，使用区域提议网络(Region Proposal Network，RPN)来提升目标的判别能力以及目标框的回归能力。之后为了应对基于孪生网络的目标跟踪无法有效利用深度网络特征的问题，Li 等[5]提出了 SiamRPN＋＋，其成功地训练了基于 ResNet(Residual Network)架构的孪生跟踪器，显著提高了跟踪精度；针对跟踪过程中非语义信息的干扰和模型更新不及时等问题，Zhu 等[6]提出了 DASiamRPN(Distractor-Aware SiamRPN)，将 SiamRPN 改进为适用于长时间的目标跟踪任务。随着孪生网络的发展壮大，SiamMask[7]结合了目标跟踪网络与目标分割网络，可以同时完成视频跟踪和实例级分割的任务。

SiamRPN 在跟踪过程中不能有效应对非语义信息的干扰以及 RPN 中特征利用不充分等问题，使其在跟踪过程中应对遮挡等情况的效果不佳，复杂背景下的目标判别能力较弱[8-9]。针对以上问题，本文将原有的仅使用第一帧为固定模板改变为动态模板[10]，优化复杂环境下的相似度匹配；将目标跟踪的缓冲区应用于孪生网络的模板分支，通过稀疏的缓存目标外观来应对跟踪过程中的遮挡问题；在 RPN 的分类分支上，应用三元组损失函数，弥补特征利用不充分的不足。本算法可以将稀疏特征聚合后再通过高效的损失函数进行判别，能够增强目标跟踪的鲁棒性和精确度，从而进一步提升目标的跟踪性能。

2　所提算法

2.1　模型构建

跟踪模型的总体结构如图 1 所示，其中 k 为锚点框的个数。该结构展示跟踪模型主体部分的内容，由左至右，第一部分是结合目标缓冲区的孪生卷积神经网络，上分支为模板分支，下分支为检测分支；第二部分是RPN模块，上分支为分类分支，下

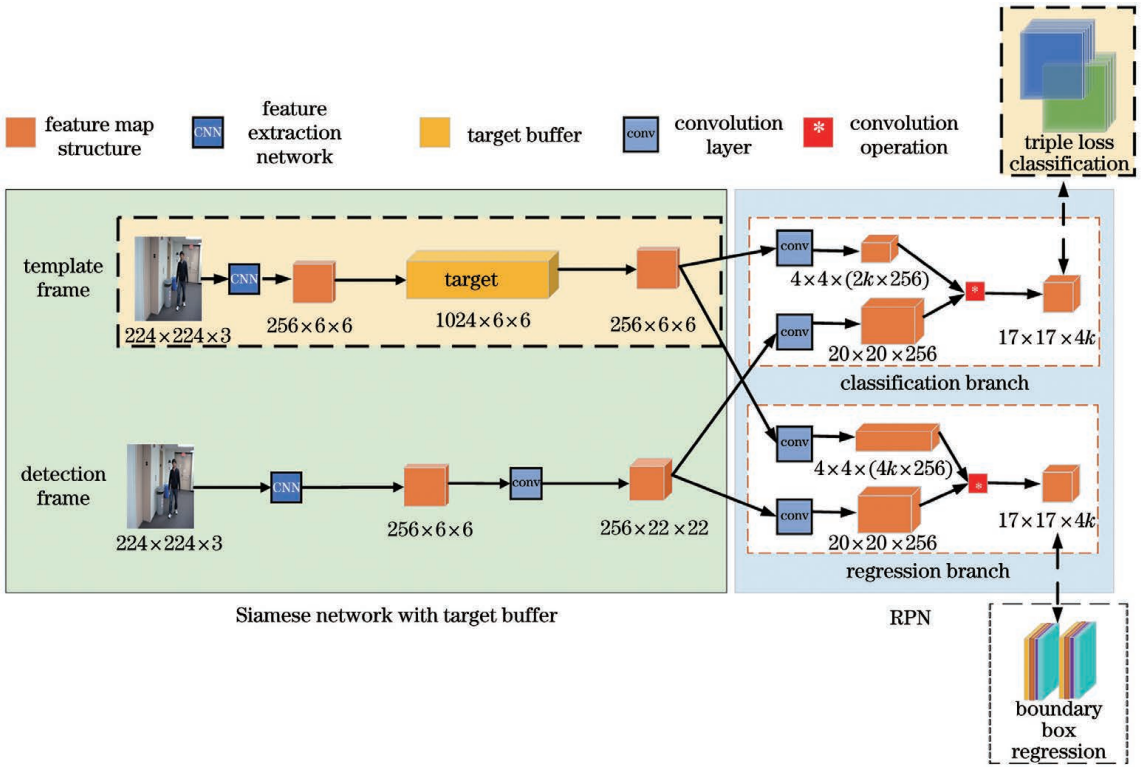

图 1　跟踪模型的总体结构

Fig. 1　Overall structure of tracking model

分支为回归分支，三元组损失分类和基于锚点的边界框回归分别作用于 RPN 的上、下两分支。基于图像序列通过端到端的方式对网络进行训练，跟踪结果利用测试集进行评估。具体改进如下：首先对孪生网络分支进行改进，将原有的 SiamRPN 的模板分支舍弃，改为结合目标缓冲区的动态模板分支，通过稀疏激活和动态缓存模板特征可以使不同帧之间的信息有效结合；然后将孪生网络双分支输出的特征应用在 RPN 中，其中分类分支和回归分支利用锚点机制来搜索卷积后特征图内的目标区域，分类分支中在不增加特征样本的前提下将原有的逻辑损失舍弃，改为三元组损失，从而提高跟踪过程中目标的表征和判别能力，回归分支则是完成对目标尺度的精确估计。

2.2 结合目标缓冲区的动态模板分支

由于目标跟踪需要对下一时刻的运动状态进行分析和预测，所以模型的更新与样本信息的保留直接影响了目标跟踪的效果。如果在相似性判别的过程中仅使用第一帧作为模板帧的方式，那么在目标发生明显形变与遮挡等复杂的情况下会造成跟踪性能不具有鲁棒性。但采用 MD-NET（Multi-Domain Convolutional Neural Networks）[11] 等在线微调的模型来更新跟踪策略或是采用光流法[12-14]进行重检测，这都会严重影响跟踪速率。所以，为了高效提升复杂环境下的跟踪性能，本文利用动态稀疏更新的模板将上一帧的跟踪结果应用到下一轮的跟踪预测过程中，并结合具有特征聚合作用的模板缓冲区来提升目标的跟踪性能。

选取尺寸为 224×224×3 的图像作为双分支的输入，特征提取网络采用的是 AlexNet 卷积神经网络[15]，经过前 5 个卷积层的作用可以得到尺寸为 256×6×6 的特征结构，稀疏激活的模板分支在缓冲区的作用下缓存目标特征，使得模板作为先验信息而得到充分利用，得到的结果与检测结果共同作为 RPN 的输入。特征缓冲区是将提取到的特征按通道进行压缩，而缓存的多个样本具有平移不变性，当输入进行少量平移时，池化层能够使输入的近似表示不变。所以，结合延时缓冲特性，此缓冲区的功能就是不断保存遮挡等复杂环境下的目标状态。

模板分支在特定的时间间隔 τ 内稀疏激活第 Z 帧图像，并在缓冲区内缓存目标特征，结合目标缓冲区的孪生网络，则孪生网络的相似性评判函数可表示为

$$f(\boldsymbol{x}_i,\boldsymbol{e}_i)=g_{\{\mathrm{clc,reg}\}}[\varphi(\boldsymbol{x}_i),\varphi(\boldsymbol{e}_i)], \tag{1}$$

$$\boldsymbol{E}=[Z_1\ Z_2\ \cdots\ Z_{\max}], \tag{2}$$

式中：i 表示跟踪过程中视频帧序列数；$\boldsymbol{x}_i\in\boldsymbol{X}$ 表示搜索分支输入的图像序列；$\boldsymbol{E}$ 表示缓冲区内的模板，$\boldsymbol{e}_i\in\boldsymbol{E}$，即一个 2D 平均池化器，模板是以张量的形式进行聚合后压缩；$g_{\{\mathrm{clc,reg}\}}$ 表示分类与回归的模板匹配与目标验证函数，使用卷积操作来判断模板帧和检测帧的相似性；φ 表示 Alexnet 特征提取器。由 BOBBY2（Buffer Based Robust High-Speed Object Tracking）[16] 的理论研究表明，缓存的目标数量不宜过多，数量过多会造成误差累积，从而直接导致目标跟踪失败，数量不足则会导致无法聚合有效特征。因此本文选取三个样本进行缓存，每经过一个时间间隔 τ，稀疏激活的模板分支就会输入第 Z 帧模板缓冲区中。

2.3 结合三元组损失的目标定位

样本特征的稀疏聚合在增强目标判别能力的同时，正负样本的特征数量也在增加，使用传统的逻辑损失可能会提取不到确切的目标表征信息。考虑到 RPN 中的分类分支需要准确区分复杂环境下的前景和背景，所以将三元组损失替代逻辑损失并引入跟踪算法中，用来解决训练过程中的正负样本不平衡，通过组合的两两正负样本，能够引导网络学习更具有判别性的特征。

逻辑损失可表示为

$$L_1(Y,V)=\sum_{\boldsymbol{x}_i\in\boldsymbol{X}}w_i\ln[1+\exp(-y\cdot v_i)], \tag{3}$$

式中：Y、V 和 $\boldsymbol{X}$ 分别表示真值标签、相似度评分和真实输入；$y\in\{-1,1\}$ 表示一个单一样本对 $(\boldsymbol{e}_i,\boldsymbol{x}_i)$ 的真值标签；v_i 表示 $(\boldsymbol{e}_i,\boldsymbol{x}_i)$ 的相似度分数，即 $v_i=f(\boldsymbol{e}_i,\boldsymbol{x}_i)$；$w_i$ 表示样本 $\boldsymbol{x}_i$ 的权重，$\sum_{\boldsymbol{x}_i\in\boldsymbol{X}}w_i=1$，$w_i>0$。在一般的逻辑损失中，平衡权重的表达式为

$$w_i=\begin{cases}\dfrac{1}{2M}, y_i\in 1\\ \dfrac{1}{2N}, y_i\in -1\end{cases}, \tag{4}$$

式中：M 和 N 分别表示正样本集 $\boldsymbol{X}_{\mathrm{p}}$ 和负样本集 $\boldsymbol{X}_{\mathrm{n}}$ 的数目，$M=|\boldsymbol{X}_{\mathrm{p}}|$，$N=|\boldsymbol{X}_{\mathrm{n}}|$。综上可得

$$L_1=-\frac{1}{MN}\sum_{i}^{M}\sum_{j}^{N}\frac{1}{2}\{\ln[1+\exp(v_{\mathrm{p},i})]+\ln[1+\exp(v_{\mathrm{n},j})]\}, \tag{5}$$

式中：v_{p} 和 v_{n} 分别表示正样本和负样本的相似性评分。由(1)式和(5)式可知，逻辑损失只利用成对损失，忽略了正样本和负样本之间的潜在关系。三

元组损失在区分正样本和负样本的基础上，挖掘两者之间的潜在关系，逻辑损失只包含总数为正负样本之和的变化损失，三元组损失则包含总数为正负样本之积的变化损失，损失函数中的变量越多，表示功能就越强大。将三元组损失函数应用在 RPN 中，使用锚点框和真实值的交并比（Intersection-over-Union，IoU）来判断特征是否属于正负样本，样本对的相似得分集 V 也可以拆分为正分集 V_{p} 和负分集 V_{n}。定义分数对的三元组损失，使用 Softmax 函数并应用匹配概率来测量每个分数对，匹配概率的公式为

$$t_{\mathrm{l}}(v_{\mathrm{p},i},v_{\mathrm{n},j})=\frac{\exp(v_{\mathrm{p},i})}{\exp(v_{\mathrm{p},i})+\exp(v_{\mathrm{n},j})}。\tag{6}$$

为了使所有分数对之间的联合概率最大化，对所有概率进行乘积，利用其负对数可以得到损失公式，即

$$L_{\mathrm{cls}}(V_{\mathrm{p}},V_{\mathrm{n}})=-\frac{1}{MN}\sum_{i}^{M}\sum_{j}^{N}\ln[t_{\mathrm{l}}(v_{\mathrm{p},i},v_{\mathrm{n},j})]=-\frac{1}{MN}\sum_{i}^{M}\sum_{j}^{N}\ln[1+\exp(v_{\mathrm{n},j}-v_{\mathrm{p},i})],\tag{7}$$

式中：$1/MN$ 表示平衡权重，用于对不同数量的样本集保持相同比例的损失。与原始的成对逻辑损失相比，三元组损失能够捕获更多的底层信息，从而在训练过程中实现更强大的表示。

在反向传播阶段，梯度的变化可直观反映出损失函数的特性，省略对应的系数项与加权项后，逻辑损失 T_{l} 的梯度可表示为

$$\begin{cases}\dfrac{\partial T_{\mathrm{l}}}{\partial v_{\mathrm{p}}}=-\dfrac{1}{2[1+\exp(v_{\mathrm{p}})]}\\[2ex]\dfrac{\partial T_{\mathrm{l}}}{\partial v_{\mathrm{n}}}=\dfrac{1}{2[1+\exp(-v_{\mathrm{n}})]}\end{cases}。\tag{8}$$

三元组损失 T_{t} 的梯度可表示为

$$\begin{cases}\dfrac{\partial T_{\mathrm{t}}}{\partial v_{\mathrm{p}}}=-\dfrac{1}{1+\exp(v_{\mathrm{p}}-v_{\mathrm{n}})}\\[2ex]\dfrac{\partial T_{\mathrm{t}}}{\partial v_{\mathrm{n}}}=\dfrac{1}{1+\exp(v_{\mathrm{p}}-v_{\mathrm{n}})}\end{cases}。\tag{9}$$

由(8)式和(9)式可以看到，$\dfrac{\partial T_{\mathrm{l}}}{\partial v_{\mathrm{p}}}$和$\dfrac{\partial T_{\mathrm{l}}}{\partial v_{\mathrm{n}}}$分别只依赖于 v_{p} 和 v_{n}，而三元组损失的$\dfrac{\partial T_{\mathrm{t}}}{\partial v_{\mathrm{p}}}$和$\dfrac{\partial T_{\mathrm{t}}}{\partial v_{\mathrm{n}}}$同时考虑 v_{p} 和 v_{n}，所以充分利用 v_{p} 和 v_{n} 所提供的信息并应用在区域候选网络的分类分支中，可以实现更加高效地判别目标。

2.4　结合 RPN 的目标跟踪

2.4.1　目标框的训练

在孪生双分支网络特征提取与模板分支特征缓存后，利用 RPN 进行目标定位和边界框回归。在目标框选取前，首先在检测分支的输入特征图上的每一个特征点生成 k 个初始边框，RPN 的双分支结构利用提取好的特征进行训练，表达式为

$$\begin{cases}\boldsymbol{A}_{w\times h\times 2k}^{\mathrm{cls}}=[\varphi(\boldsymbol{D})]_{\mathrm{cls}}*[\varphi(\boldsymbol{E})]_{\mathrm{cls}}\\\boldsymbol{A}_{w\times h\times 4k}^{\mathrm{reg}}=[\varphi(\boldsymbol{D})]_{\mathrm{reg}}*[\varphi(\boldsymbol{E})]_{\mathrm{reg}}\end{cases},\tag{10}$$

式中：$\boldsymbol{D}$ 表示检测帧的特征；$[\varphi(\boldsymbol{E})]_{\mathrm{cls}}$ 和$[\varphi(\boldsymbol{E})]_{\mathrm{reg}}$分别表示模板帧在分类分支和回归分支上的特征映射，用作卷积核；$*$ 表示卷积运算符号；$\boldsymbol{A}_{w\times h\times 2k}^{\mathrm{cls}}$ 表示分类的卷积响应图，w 和 h 分别表示响应图的宽和高，响应图中的点$(\hat{\boldsymbol{w}},\hat{\boldsymbol{h}})$包含 $2k$ 个通道向量，该向量表示原始图上相应位置处每个锚的负激活和正激活；$\boldsymbol{A}_{w\times h\times 4k}^{\mathrm{reg}}$ 表示回归的卷积响应图，响应图中的点$(\hat{\boldsymbol{w}},\hat{\boldsymbol{h}})$包含 $4k$ 个信道向量，该 $4k$ 个信道向量分别表示$[\mathrm{d}x\ \mathrm{d}y\ \mathrm{d}w\ \mathrm{d}h]$，即 4 个测量锚与真值之间的距离。分类损失采用三元组损失，回归损失采用带归一化坐标的平滑 L_1 范数损失。

在回归分支中，为了能够有效选取目标框，需要对边框进行归一化操作，其中 A_x、A_y、A_h 和 A_w 分别表示锚点框的中心点坐标和高宽，T_x、T_y、T_h 和 T_w 表示对应的真值，则归一化值为

$$\begin{cases}\delta[0]=\dfrac{T_x-A_x}{A_w}\\[2ex]\delta[1]=\dfrac{T_y-A_y}{A_h}\\[2ex]\delta[2]=\ln[T_w/A_w]\\\delta[3]=\ln[T_h/A_h]\end{cases}。\tag{11}$$

将归一化之后的值代入平滑 L_1 范数损失函数中，计算最小绝对值偏差后进行边界框回归，平滑 L_1 范数损失函数可表示为

$$L_{\mathrm{smooth-L_1}}(\delta[u],\sigma)=\begin{cases}0.5\sigma^2(\delta[i])^2,|\delta[u]|<\dfrac{1}{\sigma^2}\\[2ex]|\delta[u]|<\dfrac{1}{2\sigma^2},|\delta[u]|\geqslant\dfrac{1}{\sigma^2}\end{cases},\tag{12}$$

$$L_{\mathrm{reg}}=\sum_{u=0}^{3}L_{\mathrm{smooth-L_1}}(\delta[u],\sigma),\tag{13}$$

式中：σ 表示常数；u 表示索引号，$u=\{0,1,2,3\}$。

将三元组损失与平滑 L_1 范数损失进行加权，即可得到总的损失函数，可表示为

$$L_{\text{loss}}=L_{\text{cls}}+\lambda L_{\text{reg}}, \tag{14}$$

式中：λ 表示平衡分类损失与回归损失的超参数。通过最终的损失函数可以训练得到合适的参数，从而提取预候选框，并在目标跟踪过程中完成目标框的选取。

2.4.2　目标框的选取

二阶段检测[17]可以获得更高的准确率，使用 RPN 可以得到修正后的候选框以及每个框的类别得分。但由于跟踪所具有的特性，最高评分可能会造成误差，所以需要找到与上一帧相似的结果作为目标。本文通过目标框的大小、位置和长宽比对最终目标定位进行惩罚，可以得到与上一帧的相似度 θ，s_{c} 表示神经网络输出的目标框得分，两者相乘可以得到最终的置信得分 p_{s}，表达式为

$$p_{\text{s}}=\theta\times s_{\text{c}}。 \tag{15}$$

因为上一帧与当前帧具有最为接近的性质，所以若与上一帧相差较大，则相似度最低，需要赋予较低的权值，相似度可表示为

$$\theta=\exp\left[-\left(s_{\text{b}}\times r_{\text{b}}\times b_{\text{b}}-1\right)\right], \tag{16}$$

式中：s_{b}、r_{b} 和 b_{b} 分别表示候选框的大小、长宽比以及位置的影响因子。假设上一帧得到的跟踪结果的目标框尺寸为 (w',h')，当前帧获取的候选框尺寸为 $\{w'_{u'},h'_{u'}\}_{u'=1}^{n}$，$u'$ 为候选框的序号，n 为候选框的个数，则令

$$\begin{cases} s_{\text{b}}=c_{\text{h}}\left(\dfrac{\sqrt{w'_{u'}h'_{u'}}}{\sqrt{w'h'}}\right) \\ r_{\text{b}}=c_{\text{h}}\left(\dfrac{w'/h'}{w'_{u'}/h'_{u'}}\right) \\ b_{\text{b}}=\sqrt{\dfrac{(w'_{u'}-w')^2+(h'_{u'}-h')^2}{2L^2}} \end{cases}, \tag{17}$$

式中：$c_{\text{h}}=\max\left[x,1/x\right]$，取值越大，表示结果越不可靠；$L$ 表示搜索区域的边长。经过计算便可得到最终目标的定位结果。

3　分析与讨论

3.1　实验与测试

实验是在 45 个 Intel Xeon(R) Gold 5118 CPU @2.30 GHz 与 NVIDIA Quadro P6000 24 GB 专业图形显卡上进行的，使用深度学习框架 PyTorch1.2.0 开发环境，编程语言的版本为 3.6.5。

本文所使用的训练集为 ILSVRC2015-VID 数据集[18]和 Youtube-BB 数据集[19]。训练阶段，采用与 SiamRPN 相同的离线训练方式从 ILSVRC 数据集中随机选取样本对，从 Youtube-BB 数据集中连续选取样本对，并且在训练过程中动态生成和提取负样本，从同一视频图像的两帧中提取模板分支和检测分支的输入，使用随机梯度下降(SGD)对其进行端到端的训练。

为了证明所提算法的有效性，在 OTB100 数据集[20]上分别选取若干个具有挑战的图像序列进行测试。为了进一步体现所提算法提升目标的跟踪效果，选取 4 种具有代表性的目标跟踪算法进行测试，分别为 BOBBY2、SiamFC、SiamFC-tri(Triplet Loss in Siamese Network)和 SiamRPN，并对实验结果分别进行定性与定量分析。其中定性分析是观察跟踪序列的实际跟踪效果并进行对比评判，定量分析是选取中心位置误差、重叠率和成功率曲线下面积(AUC)三个指标来评价跟踪结果。

3.2　定性分析与讨论

OTB100 数据集中包含 100 个视频序列，共有 58897 帧视频图像，视频的采集过程中综合考虑了光照、遮挡、模糊和快速移动等不同因素的影响。所有跟踪算法的实验结果都保证公平对比。选取 4 个具有遮挡等挑战的视频序列进行测试，视频序列名称及视频属性如表 1 所示。

表 1　视频的名称及属性

Table 1　Names and properties of videos

Video name	Video attributes
Lemming	IV, SV, OCC, FM, OPR, OV
Basketball	IV, OCC, DEF, OPR, BC
Girl2	SV, OCC, DEF, MB, OPR
Soccer	IV, SV, OCC, MB, FM, IPR, OPR, BC

表 1 中 IV 表示光照变化，OCC 表示目标模糊，DEF 表示目标形变，OPR 表示平面外旋转，BC 表示复杂背景，SV 表示尺度变化，FM 表示快速移动，IPR 表示平面内旋转，OV 表示超出视野，MB 表示运动模糊，LR 表示低分辨率。5 个算法在 4 个测试视频序列上的跟踪结果如图 2 所示。

在 Basketball 序列中，目标在跟踪过程中面临着形变与遮挡等情况，在第 20 帧图像中，尽管受到短暂的遮挡，所提算法并未发生漂移，其他算法均略有漂移；在第 202 帧图像中，尽管目标发生形变，所提算法仍能很好地跟踪目标；在后续帧的图像中，由于目标的快速移动以及相似物体的干扰，其他算法明显漂移，但所提算法仍能准确跟踪。

图 2　不同视频的跟踪结果。(a)Basketball 序列;(b)Soccer 序列;(c)Lemming 序列;(d)Girl2 序列

Fig. 2　Tacking results of different videos. (a) Basketball sequence; (b) Soccer sequence; (c) Lemming sequence; (d) Girl2 sequence

在 Soccer 序列中,目标严重遮挡,而且图像序列模糊,在第 34 帧图像中,所有算法能够很好地跟踪目标,但是在第 62 帧、第 203 帧和第 318 帧图像中,其他算法均发生细小漂移,而所提算法仍能有效跟踪目标,并未出现明显漂移的现象。

在 Lemming 序列中,第 361 帧图像中存在部分遮挡,除了 SiamFC 算法发生明显漂移以外,其他算法均发生细小漂移,但所提算法没有受到影响;在第 945 帧和第 1154 帧图像中,所提算法在目标快速移动与平面内旋转等情况下仍能不受影响。

在 Girl2 序列中,第 115 帧图像中的目标受到严重遮挡,由于所提算法具有目标缓冲区和高效判别机制,在第 135 帧图像中,所提算法也没有发生漂移。

综上所述,所提算法能够在背景干扰、快速移动、模糊运动和遮挡等复杂情况下实现目标跟踪。

3.3　定量分析与讨论

3.3.1　中心位置误差测评

中心位置误差的计算公式为

$$d=\sqrt{\left(\left|\boldsymbol{O}-\boldsymbol{O}_{\mathrm{t}}\right|^{2}\right)}, \tag{18}$$

式中:$\boldsymbol{O}$ 表示由算法计算的目标中心位置;$\boldsymbol{O}_{\mathrm{t}}$ 为数据集人工标定的目标实际中心位置。中心位置误差的单位为 pixel,表示两个中心点之间像素点的个数,中心位置误差越小表示算法越好。图 3 为 4 个视频序列的中心位置误差曲线。

为了直观分析算法的性能,将所得结果取均值

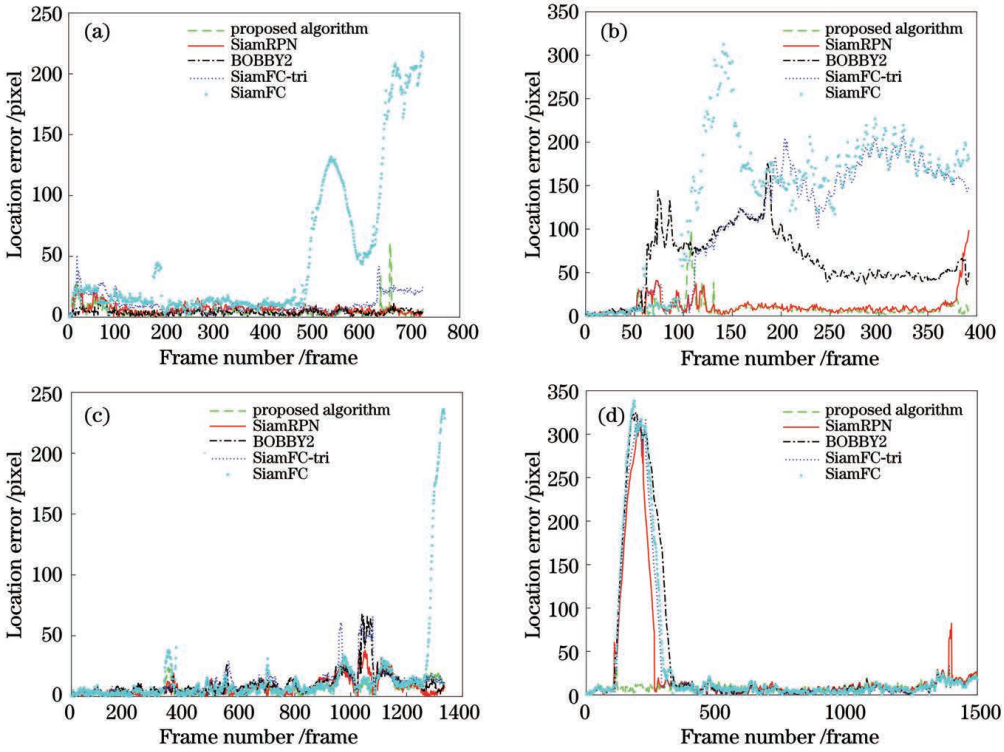

图 3　4 个视频序列的中心位置误差曲线。(a) Basketball 序列;(b) Soccer 序列;(c) Lemming 序列;(d) Girl2 序列

Fig. 3　Center position error curves of 4 video sequences. (a) Basketball sequence; (b) Soccer sequence; (c) Lemming sequence; (d) Girl2 sequence

后进行对比,结果如表 2 所示。

表 2　不同算法的平均中心位置误差

Table 2　Mean center position error of different algorithms

Algorithm	Mean center position error/pixel			
	Lemming	Basketball	Girl2	Soccer
Proposed algorithm	11.09	12.11	10.01	19.75
SiamRPN	13.23	15.01	91.55	23.55
BOBBY2	23.15	20.11	95.15	65.59
SiamFC-tri	25.65	20.40	95.75	110.75
SiamFC	29.33	35.50	101.11	151.79

从图 3 和表 2 可以看到,所提算法在中心位置误差方面达到了最优效果,另外选取测试集中 50 个序列进行计算,相较于 SiamRPN 算法,所提算法的平均中心位置误差降低了 25.56 pixel。

3.3.2　重叠率测评

重叠率的计算公式为

$$o=\frac{A(\bar{S}\cap S)}{A(\bar{S}\cup S)},\tag{19}$$

式中:$\bar{S}$ 表示由算法估计的目标覆盖范围;S 表示由数据集人工标定的目标真实覆盖范围;A 表示面积。目标尺度估计的越准确,重叠率越高。图 4 为 4 个视频序列的重叠率曲线,将所得结果取均值后进行对比,结果如表 3 所示。

表 3　不同算法的平均重叠率

Table 3　Average overlap rate of different algorithms

Algorithm	Average overlap rate /%			
	Lemming	Basketball	Girl2	Soccer
Proposed algorithm	0.89	0.79	0.77	0.61
SiamRPN	0.87	0.74	0.64	0.57
BOBBY2	0.79	0.70	0.59	0.39
SiamFC-tri	0.74	0.68	0.54	0.36
SiamFC	0.73	0.47	0.51	0.25

从图 4 和表 3 可以看到,选取测试集中 50 个序列进行计算,相较于 SiamRPN 算法,所提算法的平均重叠率提高了 25.2%,说明所提算法的平均重叠率与平均中心位置误差都取得了最佳效果。

3.3.3　成功率曲线下面积

AUC是利用曲线下面积来表示成功率,所得结

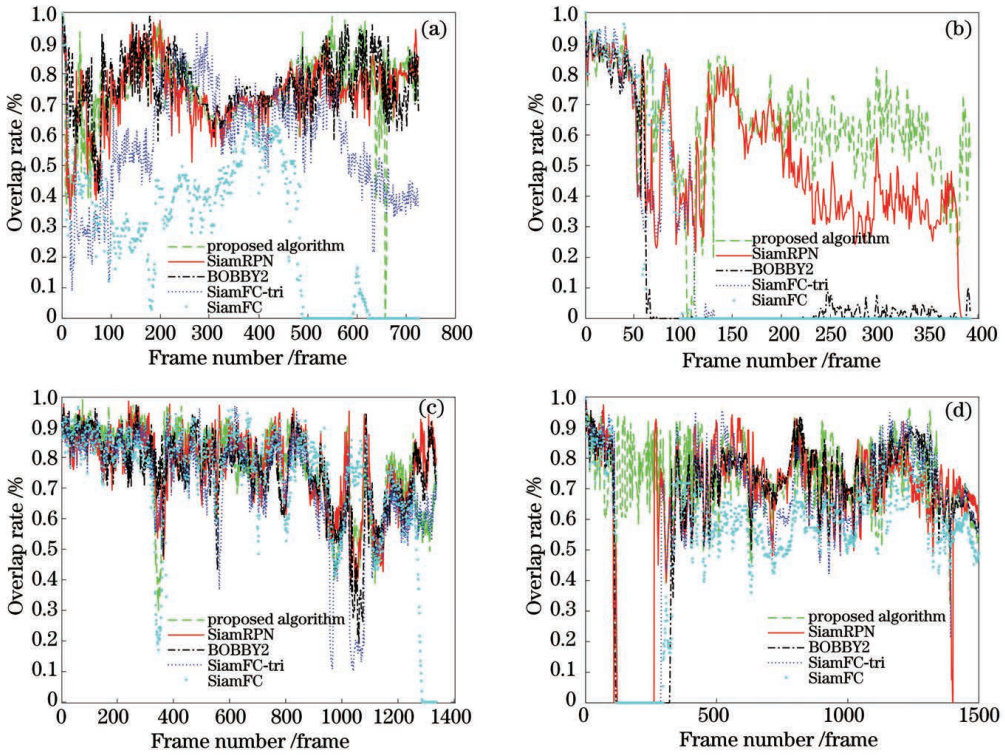

图 4　4 个视频序列的重叠率曲线。(a)Basketball 序列;(b)Soccer 序列;(c)Lemming 序列;(d)Girl2 序列

Fig. 4　Overlap rate curves of 4 video sequences. (a) Basketball sequence; (b) Soccer sequence; (c) Lemming sequence; (d) Girl2 sequence

果越大,跟踪效果越好。为了使实验结果更具有说明性,在 OTB100 测试序列上进行测试,同样也绘制了 50 个遮挡视频的成功率曲线。OTB100 数据集的测试结果如图 5 所示。

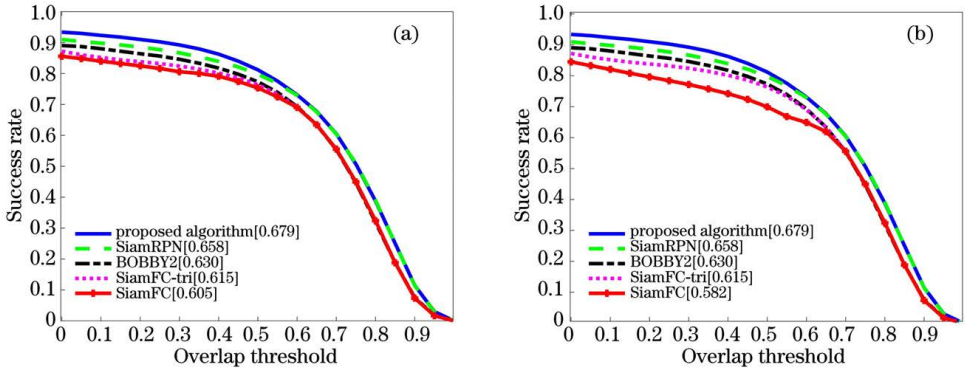

图 5　不同情况下 OTB100 数据集的测试结果。(a)整体数据集;(b)遮挡数据集

Fig. 5　Test results of OTB100 dataset under different conditions. (a) Overall dataset; (b) occlusion dataset

从图 5 可以看到,在两种测试集中,所提算法都获得了最大的 AUC 分数,均达到 0.679,相较第二名提高了 0.021,对于整体数据集来说,结果提升很大,因此证明所提算法能够有效提升目标的跟踪性能以及有效应对目标受到遮挡等复杂场景。

4　结　　论

为了解决 SiamRPN 在目标被短时遮挡以及特征判别能力不足的情况下定位不准确的问题,本文提出结合目标缓冲区以及三元组损失函数的目标跟踪算法。目标缓冲区能够在目标跟踪过程中稀疏地缓存目标特征信息,动态聚合的特征有利于在遮挡和形变等复杂场景下持续追踪目标;三元组损失替代逻辑损失,强化了正负样本之间的联系,提升了目标的判别能力。实验结果表明,所提算法在目标遮挡和形变等复杂场景下能够有效提升目标的跟踪性能,使目标定位更加准确。

参 考 文 献

[1] 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.

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

[3] Dong X P, Shen J B. Triplet loss in Siamese network for object tracking[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11217: 472-488.

[4] 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.

[5] Li B, Wu W, Wang Q, et al. SiamRPN++: evolution of Siamese visual tracking with very deep networks[C]//2019 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.

[6] 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.

[7] Wang Q, Zhang L, Bertinetto L, et al. Fast online object tracking and segmentation: a unifying approach[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 1328-1338.

[8] Wang D W, Fang H Y, Liu Y, et al. Algorithm for panoramic video tracking based on improved SiameseRPN[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241008.
王殿伟, 方浩宇, 刘颖, 等. 一种基于改进 SiameseRPN 的全景视频目标跟踪算法[J]. 激光与光电子学进展, 2020, 57(24): 241008.

[9] Zhou W, Tang H L, Li G D, et al. DDAT target tracking algorithm based on occlusion detection mechanism[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241501.
周维, 唐华龙, 李观德, 等. 基于遮挡检测机制的 DDAT 目标跟踪算法[J]. 激光与光电子学进展, 2020, 57(24): 241501.

[10] Shen Y L, Wu Z D, Zhao R J, et al. Long-term object tracking based on model updating and fast re-detection[J]. Acta Optica Sinica, 2020, 40(3): 0315002.
沈玉玲, 伍忠东, 赵汝进, 等. 基于模型更新与快速重检测的长时目标跟踪[J]. 光学学报, 2020, 40(3): 0315002.

[11] Nam H, Han B. Learning multi-domain convolutional neural networks for visual tracking[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 4293-4302.

[12] Lu H C, Lu S P, Wang D, et al. Pixel-wise spatial pyramid-based hybrid tracking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 22(9): 1365-1376.

[13] Dai K H, Wang Y H, Yan X Y. Long-term object tracking based on Siamese network[C]//2017 IEEE International Conference on Image Processing (ICIP), September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 3640-3644.

[14] Zhang J, Hao Z H, Liu J. Template-updating algorithm based on optical flow mapping in object tracking[J]. Laser & Optoelectronics Progress, 2020, 57(22): 221507.
张静, 郝志晖, 刘婧. 目标跟踪中基于光流映射的模板更新算法[J]. 激光与光电子学进展, 2020, 57(22): 221507.

[15] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[16] Lee K, Tai J J, Phang S K. BOBBY2: buffer based robust high-speed object tracking[EB/OL]. (2019-10-18)[2020-12-01]. https://arxiv.org/abs/1910.08263.

[17] 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.

[18] Real E, Shlens J, Mazzocchi S, et al. YouTube-BoundingBoxes: a large high-precision human-annotated data set for object detection in video[C]//2017 IEEE Conference on Computer Vision and

Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 7464-7473.

[19] 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.

[20] Wu Y, Lim J, Yang M H. Online object tracking: a benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2411-2418.