DOI:10.11918/201910175

一种基于改进 RT-MDNet 的全景视频目标跟踪算法

王殿伟[1],方浩宇[1],刘 颖[1],伍世虔[2],谢永军[3],宋海军[3]

(1. 西安邮电大学 通信与信息工程学院,西安 710121; 2. 武汉科技大学 信息科学与工程学院,武汉 430081;
3. 中国科学院 西安光学精密机械研究所,西安 710119)

摘 要: 为了解决全景视频目标跟踪过程中,由于光照条件变化、相似背景干扰、目标运动时产生的形变和尺度变化等因素的影响,在跟踪中会出现目标漂移、目标丢失等情况,进而导致目标跟踪算法成功率低,鲁棒性差等问题,提出一种基于长短期记忆网络和改进 Real-Time MDNet 网络的全景视频目标跟踪方法. 算法首先采用浅层卷积神经网络提取特征,并利用自适应的 RoIAlign 减少特征提取过程中的像素损耗,而后运用目标特征在线更新最后一个全连接层的权重,在全连接层中实现前景背景分离并提取出目标区域,然后通过长短期记忆网络自适应地选取目标框的尺度,最终输出目标位置信息. 实验结果表明:单目算法应用在全景数据集时,难以适应全景中的尺度变化和背景变化,改进算法利用 3 层长短期记忆网络构建的尺度预测模块,可以有效地应对全景数据存在的尺度变化和目标形变问题,在保持较好的跟踪精度的同时,可以有效地应对目标跟踪中出现的小目标、目标遮挡、多目标交叉运动的情况,获得更好的视觉效果和更高的重叠率得分.

关键词: 目标跟踪;深度学习;全景视频;长短期记忆网络;RT-MDNet

中图分类号: TP391.41;TP183 **文献标志码:** A **文章编号:** 0367-6234(2020)10-0152-09

Improved RT-MDNet for panoramic video target tracking

WANG Dianwei[1], FANG Haoyu[1], LIU Ying[1], WU Shiqian[2], XIE Yongjun[3], SONG Haijun[3]

(1. School of Communications and Information Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;
2. School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081, China;
3. Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China)

Abstract: In the process of panoramic video target tracking, the target deformation and scale changes caused by light change, interference of similar background, and object moving may result in target drift or missing, leading to low success rate and poor robustness. To address these issues, a target tracking method based on long short-term memory (LSTM) network and improved Real-Time MDNet (RT-MDNet) network was proposed. First, shallow convolution neural network was utilized to extract features, and adaptive RoIAlign was adopted to reduce pixel loss in the convolution process. Then, the weight of the last layer of the full connection layers was updated online by utilizing the target features to achieve foreground background separation and extract the target area. Lastly, the scale of the target box was selected adaptively by means of LSTM, and the target position information was thus obtained. Experimental results show that monocular vision algorithm could hardly adapt to the scale change and background change when applied in panoramic dataset, while the proposed method that utilizes 3-layer LSTM network to construct scale prediction module could effectively solve these problems. The algorithm can efficiently deal with the situations of small target, target occlusion, and cross motion of multiple targets in target tracking while maintaining accuracy, achieving better visual effect and higher overlap rate score.

Keywords: target tracking; deep learning; panoramic video; LSTM; RT-MDNet

目标跟踪是在视频序列中给定第 1 帧目标位置信息后,能够估计之后视频帧中同一目标位置与尺度信息的算法,在智能交通系统、监控系统等领域都有广泛的应用[1]. 目标跟踪算法受相似背景干扰、目标遮挡、目标尺度变化等因素的影响,导致精度较差和适用性较差,因此,如何提高目标跟踪算法鲁棒性和准确性是一项挑战[2].

近些年来深度学习的运用,使计算机视觉领域的发展更为迅速. Nam 等[3]提出了 MDNet,使用了卷积神经网络结构,用于学习目标的通用特征表示. Yun 等[4]结合监督学习和强化学提出 ADNet,训练网络学习识别目标,通过强化学习预测目标的变化姿态及尺度,算法较好地解决了尺度变化的问题,但

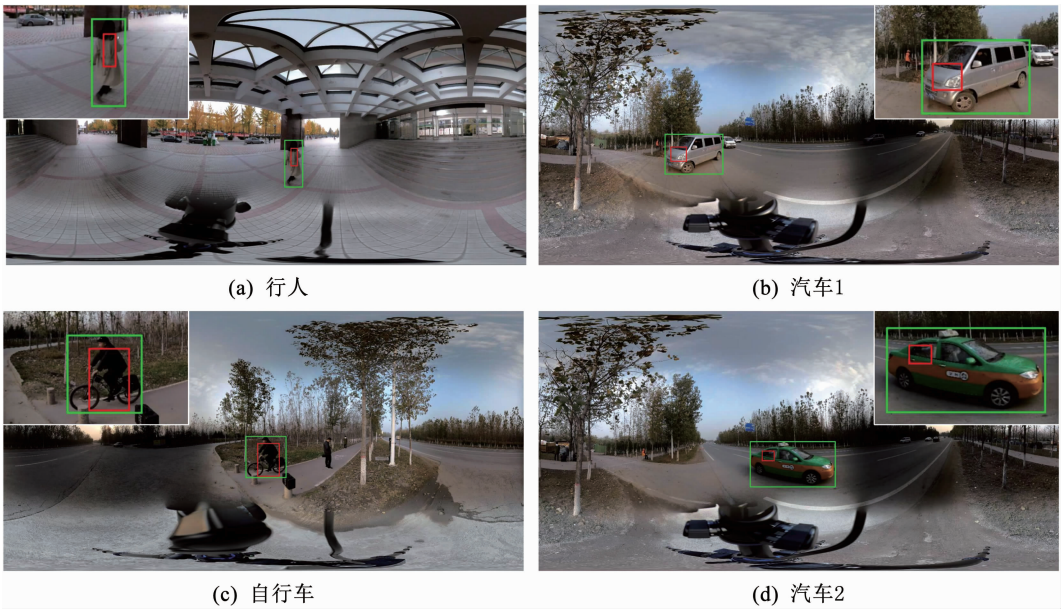
收稿日期: 2019-10-25
基金项目: 公安部科技强警基础研究专项项目(2019GABJC42);陕西省自然科学基础研究计划(创新创业"双导师")研究项目(2018JM6118);西安邮电大学研究生创新基金(CXJJLY2018033)
作者简介: 王殿伟(1978—),男,副教授,硕士生导师;
方浩宇(1994—),男,硕士研究生
通信作者: 方浩宇,fanghaoyu54057@163.com

精度不佳. Li 等[5]将 Siamese FC 与 RPN 网络相结合提出 Siamese RPN，利用相关滤波的方法提升了跟踪精度，具有实时的性能，但算法易受到背景的干扰. Jung 等[6]在 MDNet 的基础上提出 RT-MDNet，设计损失函数和采用自适应的 RoIAlign，简化特征提取网络结构，在保持了相同精度的同时，将速度提升了近 25 倍，但是算法对于目标尺度变化估计很局限，无法直接应用于全景视频图像的目标跟踪.

针对上述问题，本文提出了一种利用长短期记忆网络（Long Short-Term Memory，LSTM）改进 RT-MDNet 的目标跟踪算法，改进算法增大网络的输入以适应全景图像的输入特征，调整生成样本尺度，训练网络能更好地适应全景图像的目标形变，提高网络跟踪精度. 在原有的网络结构中增加尺度变化模块，利用 LSTM 网络学习尺度变化过程，结合之前视频帧的位置信息，自适应地调整当前视频帧的尺度变化程度，以适应全景图像中目标跟踪的尺度变化和目标形变问题. 算法很好地提高了跟踪精度，保持了一定的运算速度.

1　全景视频的目标跟踪

全景数据具有更高的分辨率，同时伴随着更复杂的场景和更高的计算要求，目标对象与摄像头相对运动时，距离的变化在跟踪中会以尺度变化的方式反映出来，当目标对象与摄像头距离越靠近，这种尺度变化程度会更严重[7]. RT-MDNet 算法对于尺度变化的映射较为简单，不能很好地适应全景视频中的变化幅度，训练 RT-MDNet 用于全景视频序列的目标跟踪时，实验结果如图 1 所示.



(a) 行人　(b) 汽车1　(c) 自行车　(d) 汽车2

图 1　不同场景下出现的尺度变化问题

Fig. 1　Scale changes in different scenarios

图 1 中为原网络输出结果与真实值，实验结果覆盖的多个场景均出现了很大程度的尺度变化，而改进前原网络应对尺度变化的能力很弱，需要分析全景图像成像方式和数据特性，针对全景数据调整网络结构，使其具有更好的适用性和应对尺度变化的能力.

1.1　本文算法流程框架

RT-MDNet 使用 BoundingBox regression 方法对边框进行调整，BoundingBox regression 根据第 1 帧真实值和预选值做线性映射改善目标尺度变化. 在全景视频中尺度随着目标的运动有规律的变化，在跟踪过程中仅使用第 1 帧做线性映射难以估计目标的尺度变化. 针对已有算法应用于全景图像目标跟踪时，跟踪精度较低且尺度变化适应性差的问题，提出了一种基于改进 RT-MDNet 的全景视频目标跟踪算法. 随着视频序列的移动，依据 LSTM 网络拥有长时间记忆单元的优势，结合不同频帧之间的尺度变化信息，利用神经网络学习数据集中尺度变化的方式，算法的整体流程如图 2 所示.

由图 2 可知输入图像经过共享的 3 个卷积层提取特征图，经过 Adaptive RoIAlign 提取出预选框特征送入全连接层区分前景背景，最后目标框经过 LSTM 网络自适应的选取目标框尺度，LSTM 网络输出最终的改进结果. 网络整体参数针对全景数据进行改进，使网络更加适用全景数据的特性，使用 Adaptive RoIAlign 进行特征提取降低了计算成本减少了卷积过程损耗，利用区域间的损失函数加强了网络对于相似目标的区分能力，提升了网络的跟踪精度.

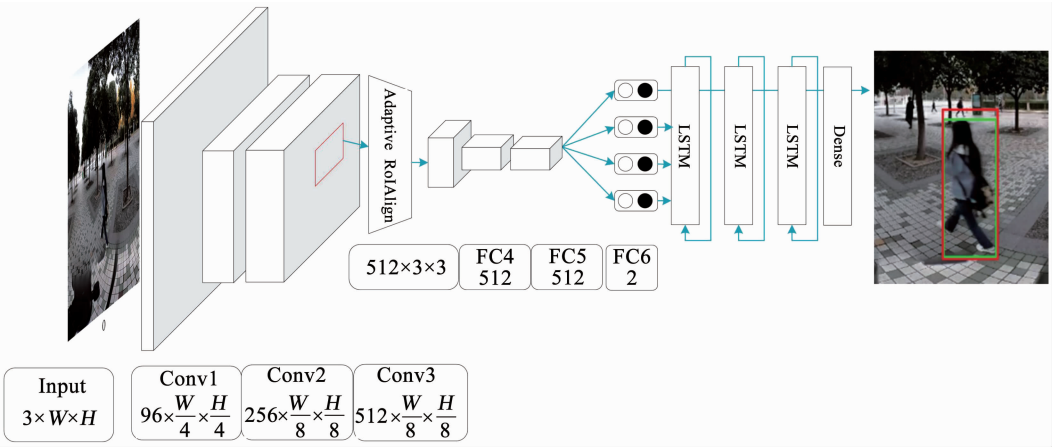


图 2 本文算法整体网络框架

Fig. 2 Network architecture of the proposed algorithm

1.2 实例间区分的损失函数

RT-MDNet 的损失函数引入了实例,在区分目标背景的同时,可较好地在特征空间中将不同视频序列的目标进行区分. RT-MDNet 的最后一个全连接层根据输入的视频序列在线调整参数,输出网络得分,并通过 Softmax 区分目标对象与背景干扰,通过另一个 Softmax 区分不同视频域之间的目标类. 整体的损失函数 L 为

$$L = L_{\text{cls}} + \alpha \cdot L_{\text{inst}}, \tag{1}$$

式中 L_{cls} 和 L_{inst} 分别为目标背景二分类和实例嵌入的损失函数,α 是控制两个损失函数之间的超参数.

每次迭代处理一个视频序列,在 $k(k=1,2,\cdots D)$ 次迭代后,用得到的批量值来更新网络,在第 k 次迭代中的序列记为 $\hat{d}(k)$,二分类损失函数由下式给出:

$$L_{\text{cls}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{2}[y_i]_{c\hat{d}(k)} \cdot \log([\sigma_{\text{cls}}(f_i^{\hat{d}(k)})]_{c\hat{d}(k)}), \tag{2}$$

式中:$y_i \in \{0,1\}^{2\times D}$ 是真实值的 one-hot 编码,对应在视频序列 d 中 c 个类别的输出为 1 或是 0. 实例间的损失函数由下式给出:

$$L_{\text{inst}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{d=1}^{D}[y_i]_{+d} \cdot \log([\sigma_{\text{inst}}(f_i^d)]_{+d}), \tag{3}$$

式中:$+d$ 为损失函数中实例嵌入的损失只由正样本给出,算法引入了当前序列的实例特征,使当前序列中的目标分数变大,其他序列目标分数变小,用以区分其他类似对象对目标的影响.

1.3 Adaptive RoIAlign

目标跟踪与目标检测中常用 RoIPooling 作为区域特征的映射方式[4,8],通过 RoIPooling 将目标预选区域通过卷积的方式映射到固定尺寸的特征图,然后进入全连接层进行分类和预选框回归操作. RoIPooling 的局限性在于,映射的过程中会出现两次量化的过程,量化的过程会损失掉一部分特征信息. 目标足够大的时候这种损失可以忽略,然而全景视频中由于其成像特性,距离稍远的目标会呈现得很小,在持续的目标跟踪中细小的误差将会持续累积,小目标出现频繁时这种损失对原有特征产生很大的影响从而导致目标丢失.

为了解决这一问题,MaskR-CNN[9] 对 RoIPooling 改进,提出了 RoIAlign,在遍历预选框时不再进行量化操作,而是通过双线性插值来得到近似特征,以实现对目标更精准地定位. RT-MDNet 采用的 Adaptive RoIAlign 方式与 MaskR-CNN 相似,双线性插值的步长由输出的 RoI feature 的大小决定,显著提高了跟踪算法的性能. RoIAlign 整体流程如图 3 所示.

图 3 中预选框经过卷积提取到的 RoI 尺度为 $W \times H$,预期经过 RoIAlign 得到的 RoI 尺度为 $W' \times H'$,$[\cdot]$ 是舍入算子通过卷积操作得到最终的输出. Adaptive RoIAlign 图层生成 7×7 的特征图,并在图层之后应用 Maxpooling 最终生成 3×3 的特征图. 在本文算法中采用 Adaptive RoIAlign 方法映射特征图,加强算法对于全景视频中小目标跟踪的鲁棒性.

1.4 LSTM

Hochreiter 等[10] 于 1997 年在 RNN 网络基础上提出 LSTM 网络. LSTM 通过引入更新门、遗忘门和输出门,同时考虑了时间序列的机制,解决了 RNN 网络中的梯度消失问题,LSTM 网络已经在目标检测,目标跟踪领域中取得了很好的成果[11]. 在跟踪中对目标框进行调整时如果只知道当前输入,所输入的信息对尺度变化的估计是不够精确的,利用 LSTM 的记忆单元连接先前的信息结合到当前任务中,可以更好地调整原始网络的输出目标框尺度.

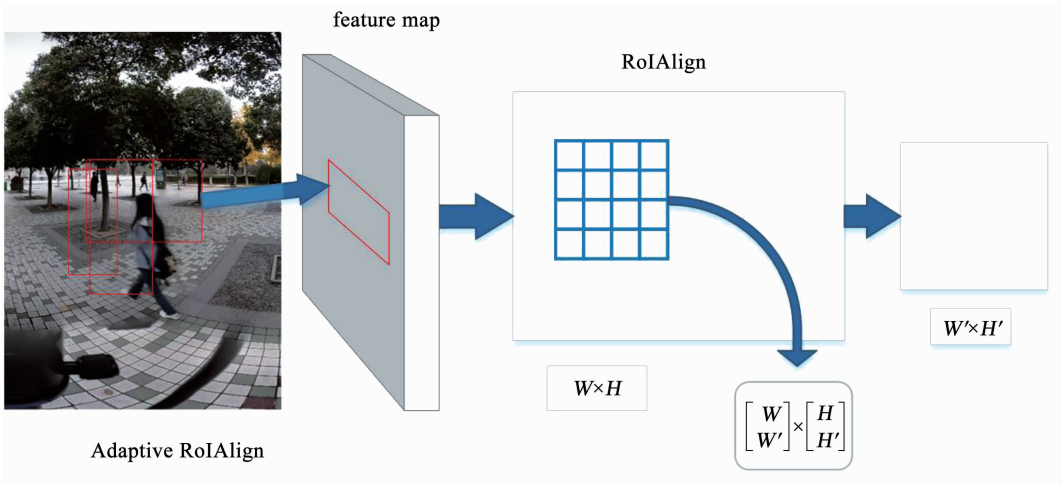


图 3　Adaptive RoIAlign 特征映射流程

Fig. 3　Feature mapping using adaptive RoIAlign

LSTM 在 t 时刻隐藏单元为：

$$\Gamma_u = \sigma(W_u[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_u), \tag{4}$$

$$\Gamma_f = \sigma(W_f[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_f), \tag{5}$$

$$\Gamma_o = \sigma(W_o[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_o). \tag{6}$$

式中：Γ_u、Γ_f 和 Γ_o 分别为更新门、遗忘门和输出门，σ 为 sigmoid 激活函数，$a^{\langle t-1\rangle}$ 为上一时刻的输出，$x^{\langle t\rangle}$ 为当前时刻的输入，W_u、W_f、W_o 和 b_u、b_f、b_o 分别是不同门的参数与偏差项. 更新门和遗忘门控制记忆细胞的更新，更新门记录当前的尺度，遗忘门选择保留更显著的特征，在记忆细胞中保留之前视频帧的尺度变化，记忆细胞公式由下式给出：

$$c^{\langle t\rangle} = \tanh(W_c[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_c), \tag{7}$$

$$c^{\langle t\rangle} = \Gamma_u * c^{\langle t\rangle} + \Gamma_f c^{\langle t-1\rangle}. \tag{8}$$

式中：$c^{\langle t\rangle}$ 是候选值，W_c 和 b_c 是参数与偏差项. 候选值 $c^{\langle t\rangle}$ 由当前时刻的输入 $x^{\langle t\rangle}$ 得到，$c\langle t\rangle$ 和前一时刻的记忆细胞 $c^{\langle t-1\rangle}$ 通过更新门和遗忘门得到当前的记忆细胞 $c^{\langle t\rangle}$，$c^{\langle t\rangle}$ 与输出门共同决定当前尺度变化的输出为

$$a^{\langle t\rangle} = \Gamma_o * \tanh c^{\langle t\rangle}, \tag{9}$$

式中：$c^{\langle t\rangle}$ 是经过输出门 Γ_o 得到当前网络的输出 $a^{\langle t\rangle}$. 本文设计的网络结构由 3 层 LSTM 和 1 个全连接层组成，整体的预测网络结构如图 4 所示.

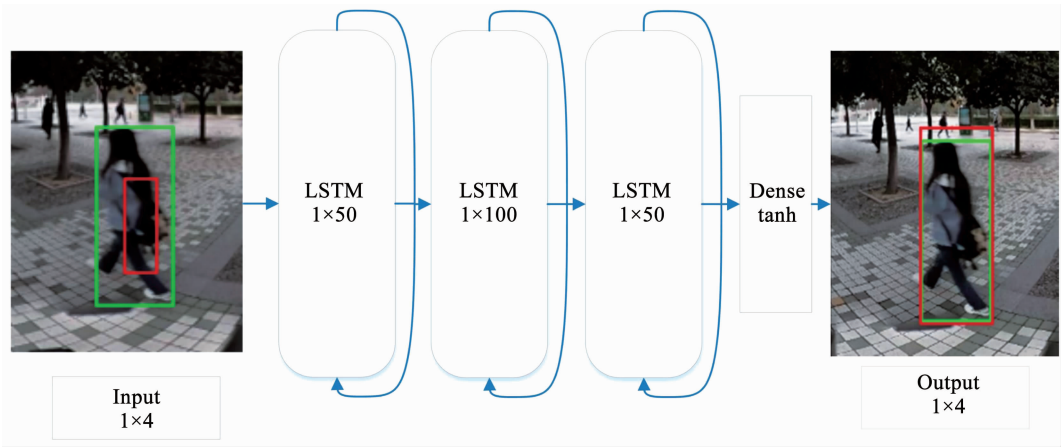


图 4　LSTM 网络结构

Fig. 4　LSTM network structure

目标在全景视频中的尺度变化方式受其位置因素的影响，变化模式较为单一. 深层次的 LSTM 网络利用多层的神经网络从多个不同维度理解尺度特征的变化，在多个层次中分解输入尺度特征，低维度输入映射到高维度相当于将低维特征分解到多个维度，再利用高维度的特征拟合全景视频尺度变化方式，在高维空间中学习运动规律，更容易学习并且能达到更高的准确率. 随着视频帧的进行，LSTM 学习在不同时刻多维度的尺度表达并将其特征保留在记忆细胞中，从高维度学习解决尺度变化的问题.

神经网络中增加网络层数可以拟合更加复杂的映射，因此增加神经网络深度是网络搭建中有效的

优化方式. 但是过深的神经网络不仅会造成过拟合，而且会造成计算资源的浪费. 为平衡网络计算复杂度以及追踪的精度，本文设置 3 组实验来验证 LSTM 的层数选择，LSTM 分别为 2 层、3 层、4 层. 网络中使用尽可能少的神经元数量达到需求的准确率是搭建结构中的重点. 在实验中采用 Adam 算法优化网络训练，针对归一化的数据采用 tanh 激活函数，在多次实验中衡量损失值的变化趋势调整学习率和训练批量，使损失值下降的更为平滑，并且梯度向最优方向迭代. 通过实验对比网络节点数对精度的影响，本算法选择先分解输入特征再聚合的网络结构，最后通过全连接层输出目标框. 图 5 为选取一部分实验数据进行网络预训练的实验结果图.

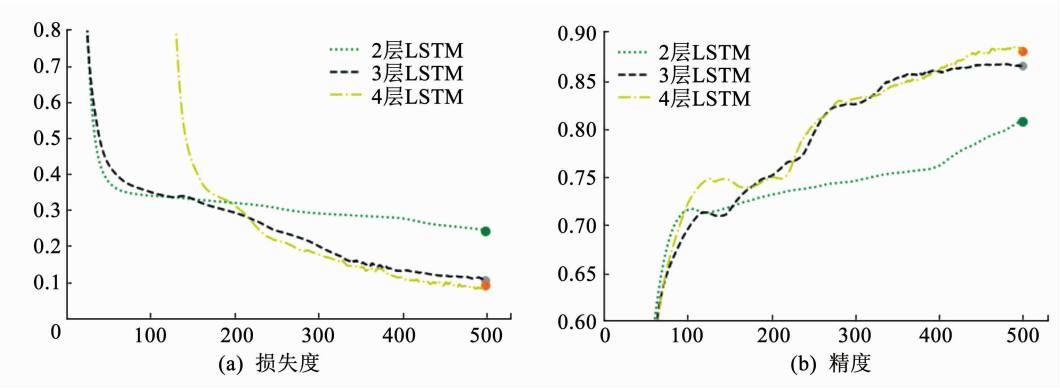


图 5 不同网络层数损失值和精度结果对比

Fig. 5 Comparison of loss values and accuracy of different network layers

图 5 中分别代表 2 层、3 层、4 层 LSTM 网络在训练中的损失值与精度的变化曲线. 3 次实验中均选择相同的实验数据和网络参数，2 层 LSTM 网络损失值局部收敛得更快一些，在精度和损失值趋于平缓时准确率并不理想. 2 层网络在训练中受深度的限制，精度提升缓慢原因在于提取的特征少，处于当前最优的情况，损失值不再下降. 3 层和 4 层的 LSTM 趋近于收敛后，可以达到更低的损失值和更高的准确率，3 层的 LSTM 在达到准确率要求的同时运用了更少的计算资源.

表 1 3 种网络结构参数量对比

Tab. 1 Comparison of parameter numbers of three network structures

网络类型	2 层 LSTM	3 层 LSTM	4 层 LSTM
参数量	31 404	101 804	122 004

原网络输出每帧目标的位置信息和尺度信息，利用 LSTM 网络的记忆特性结合之前帧的位置信息和尺度信息，学习当前帧的目标尺度变化. 输入经过 3 层 LSTM 网络得到输出 $a^{\langle t\rangle}$，$a^{\langle t\rangle}$ 再经过全连接层得到当前网络的尺度变化. 图 6 为改进后 LSTM 网络的目标框与原网络输出目标框的结果.

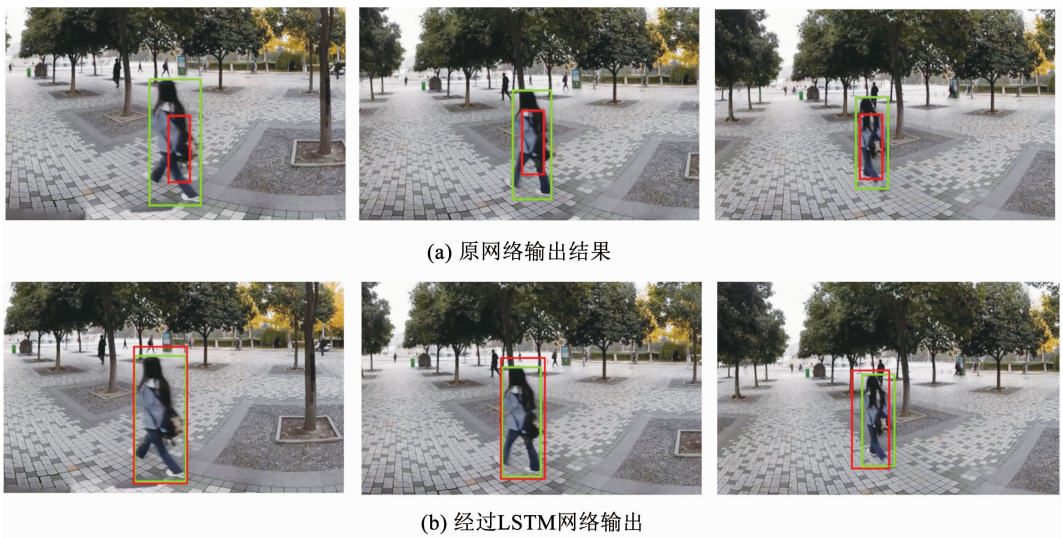


图 6 原网络与改进网络实验结果对比

Fig. 6 Comparison of experimental results of original network and improved network

图 6(a)为原网络输出结果，图 6(b)为改进网络的输出结果. 由图可知目标由近及远的过程中出现了大幅度的尺度变化问题，原网络难以适应尺度变化，经过改进的网络在跟踪中能自适应调整目标框尺度，取得更好的跟踪效果.

2　实验结果与分析

目前常用的目标跟踪算法都是基于公开数据集，如 OTB[12]，VOT[13] 等数据集，尽管在公开数据集中可以获得特征表达，但由于数据集场景还是较为单一，导致在跟踪方面的有效性受到数据集的限制. 为了在全景数据上有更佳的表现力，就需要可用于训练和测试的全景数据集. 为了解决上述问题，本文建立了用于目标跟踪的全景数据集，该数据集包含标注了多个场景、不同时间(白天、夜晚)条件下的行人、车辆等数据，可以实现神经网络端到端的训练. 所有训练及测试数据集均为泰科易 720 Pro 七目全景相机采集所得，分为 4 个类别进行了标注，处理后的图片分辨率 2 000 × 1 000.

硬件配置为 CPU Intel Xeon E5-2620v4 × 2，显卡 GPU NVIDIA Titan XP × 4. 在 Ubuntu 系统中使用 Python 作为实验平台，训练的 LSTM 网络用 Keras 框架搭建.

2.1　主观分析

为了评估算法在全景图像中的有效性，本文选取了多个不同场景不同目标的全景视频作为测试数据，并与 MDNet，ADNet，RT-MDNet 和 Siamese RPN 算法的跟踪结果做主观和客观分析. 实验结果中全景视频序列涵盖了目标变形，目标旋转，光照变化，长时间跟踪等诸多现实挑战情况，为了突出对比性能结果的好坏，对整幅全景图进行了截取，并选取出其中具有较复杂的尺度变化问题的视频序列. 结果图中不同的线型代表不同的跟踪算法中的目标框，其主观结果如图 7 所示.

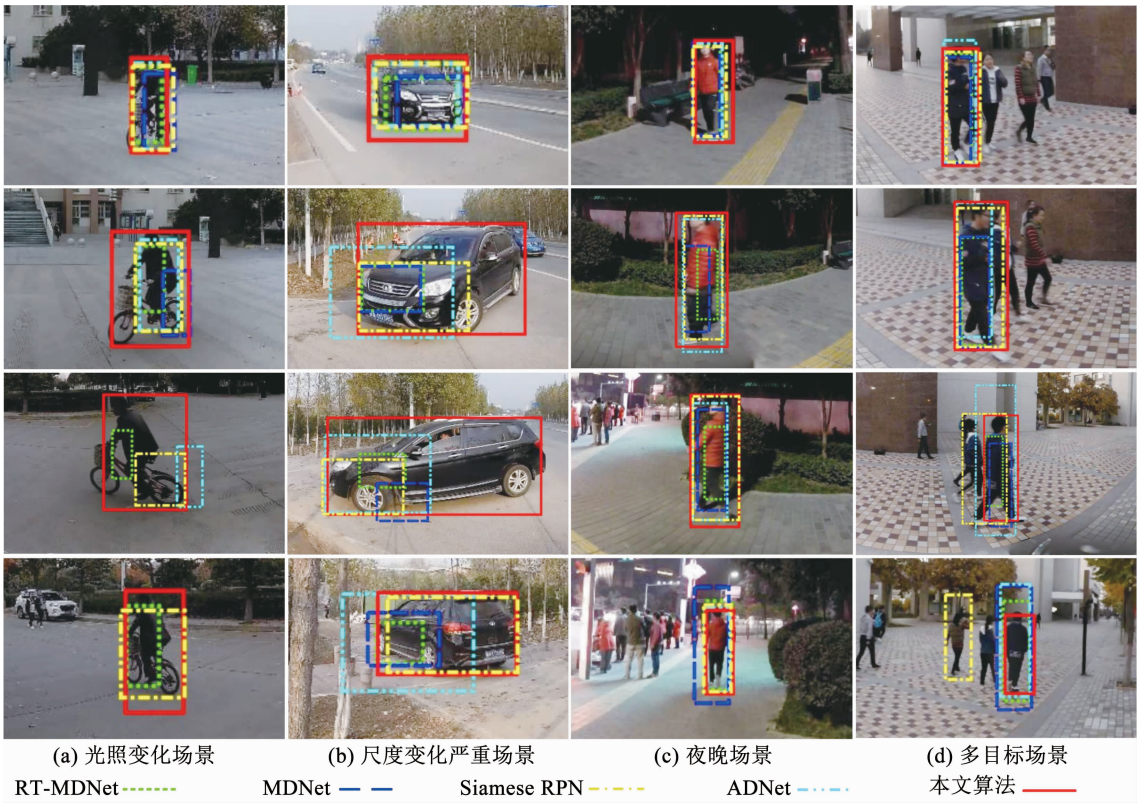


图 7　4 个不同场景下不同算法结果对比

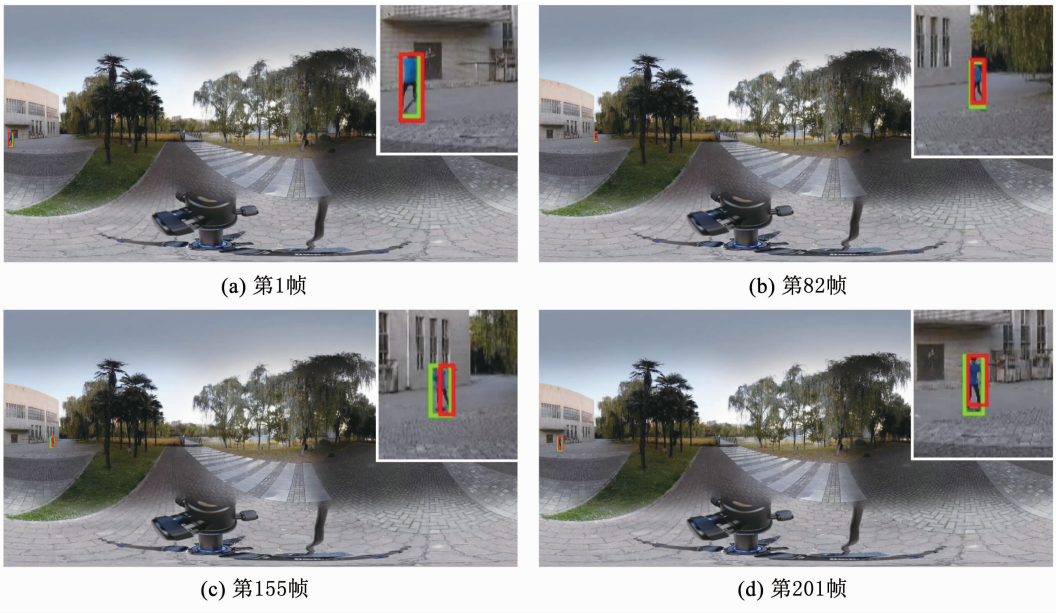
Fig. 7　Comparison of experimental results of different algorithms in four scenarios

图 7(a)至(d)分别为自行车、汽车、夜晚和白天的行人视频序列，序列中均出现了较大程度的尺度变化和外观变化. 在图 7(a)序列中目标旋转和光照的影响比较大，MDNet 和 ADNet 不能很好地应对这种变化，出现了跟踪丢失的情况，本文算法对受光照影响的目标跟踪效果较好. 图 7(b)序列中物体出现了剧烈的旋转和尺度变化，ADNet 和 Siamese RPN 具有应对尺度变化的模块，在图 7(b)中对于尺度变化的适应比 RT-MDNet 和 MDNet 稍好一些，但是在全景数据上依然很难达到很好的视觉效果，本文改进算法也能较好地适应这种情况. 图 7(c)中 5 种算法均有较好的准确率，图 7(d)中 Siamese RPN 在受到具有相似特征的背景干扰时发生了目标丢失的情况，本文改进算法在准确跟踪目标的同时，目标框能

够结合之前视频帧自适应的变化. 图 8、图 9 和图 10 为采用本文算法得到的完整实验结果与真实值对比及其跟踪目标的放大图.

由图 8 可见,全景视频序列中小目标较为普遍,小目标尺度变化程度不明显,本算法在应对全景视频中的小目标时,依然能够准确稳定地追踪,具有较好的鲁棒性.

由图 9 可见,在多个目标交叉运动时,虽然受多个相似目标的影响出现了小幅度的漂移,但在后续视频帧中仍然可以稳定跟踪目标对象. 本算法在区分相似的群目标时,能持续跟踪选定目标,具有较好的自适应跟踪能力.

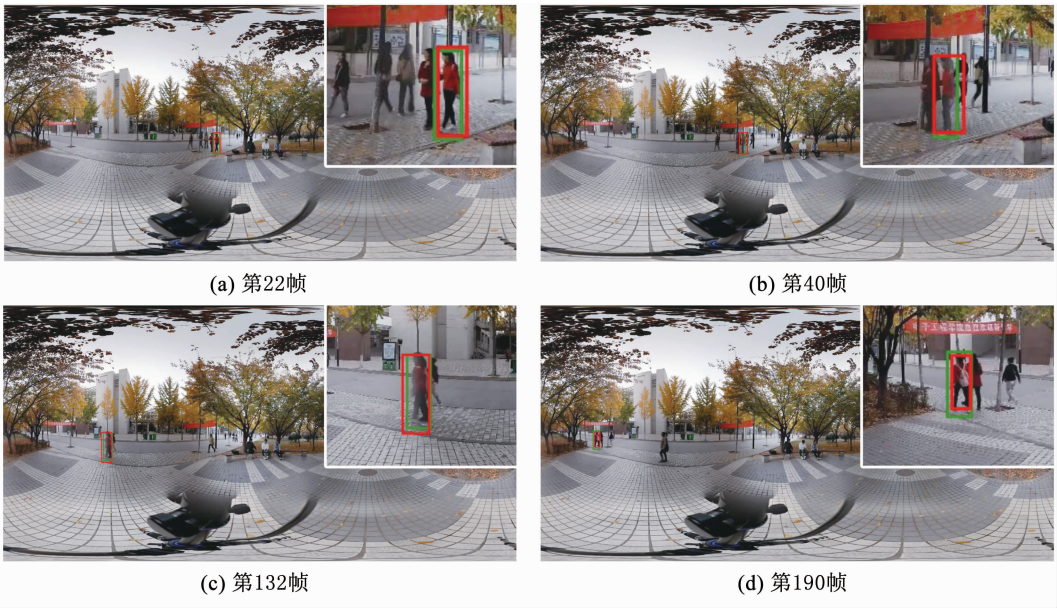


(a) 第1帧 (b) 第82帧

(c) 第155帧 (d) 第201帧

图 8 小目标情况下的实验结果

Fig. 8 Experimental results of small target in panoramic pictures



(a) 第22帧 (b) 第40帧

(c) 第132帧 (d) 第190帧

图 9 多个目标交叉运动的实验结果

Fig. 9 Experimental results of cross motion of multiple targets in panoramic pictures

图 10 中出现了目标遮挡的问题,对跟踪结果产生了一定的影响,但接下来的视频帧目标重新出现改进算法能够继续跟踪目标,本算法在应对遮挡问题上仍有不错的表现.

综上所述,RT-MDNet 与 MDNet 都达到了很好的精度,但缺少对目标尺度变化的估计. ADNet 和 Siamese RPN 具有应对尺度变化的能力,但是不能满足全景数据的目标变化. 在速度上全景图像由于具有很高的分辨率所以很难达到实时的要求,本文算法在应对不同光照条件、不同目标时可以较好地应对目标的尺度变化,并提供了准确率和重叠率.

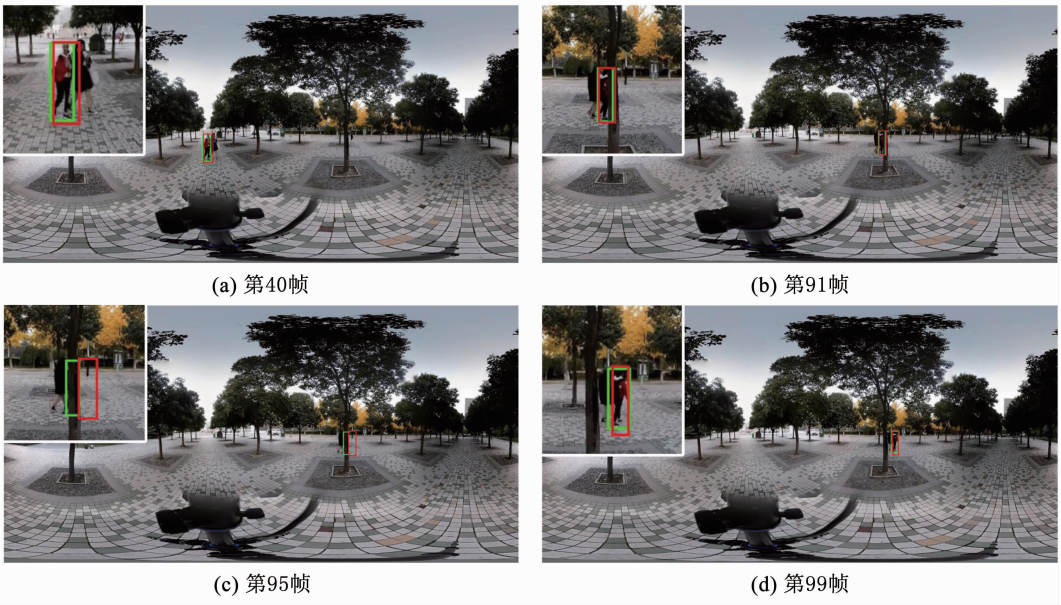


图 10　目标遮挡情况下的实验结果

Fig. 10　Experimental results of target occlusion in panoramic pictures

2.2　客观分析

为了评估算法性能，利用重叠率（Intersection over Union，IOU）和距离精度作为客观分析指标来评估算法. 重叠率表示跟踪结果与真实值重叠部分与整体之间的比值，距离精度表示跟踪结果中心位置与真实值结果中心位置的欧氏距离. 评估性能时须得到当前帧重叠率和距离精度，当大于一定阈值判定为预测准确，判定为预测准确的视频帧与整体视频帧的比率称之为成功率和精度. 在全景图像数据集上试验得到预测结果 IOU 和目标框，可视化为曲线图 11. 计算两个标准中不同阈值所对应成功率和精度来生成这两个对比图，根据其中的成功率和准确率得分对跟踪器进行排名.

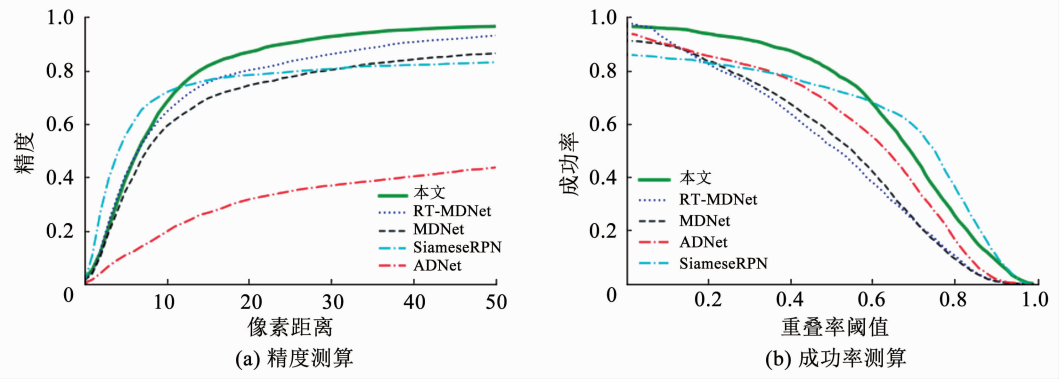


图 11　4 种算法在全景数据集上的测试结果

Fig. 11　Test results of four algorithms on panoramic dataset

图 11 给出 RT-MDNet、MDNet、ADNet 和 Siamese RPN，4 种算法与本文改进算法精确率和成功率的比较. ADNet 丢失目标的视频帧较多，所以在精度图中的表现较差，而在成功率图中 IOU 高于 RT-MDNet 和 MDNet. Siamese RPN 应对尺度变化的能力强于其他 4 种算法，但成功率略低于 MDNet 和 RT-MDNet. 从图 11 中可以看出本文改进算法在精度测算图和成功率测算图中对于原算法均有明显的提升. 表 2 中给出各算法在欧氏距离阈值为 20 像素时跟踪器的精确率，IOU 大于阈值 0.5 时跟踪器的成功率，数据集距离精度的平均值，即平均中心位置误差和基于全景数据集的平均 FPS.

由表 2 可知，由于全景图像具有较大的分辨率，复杂的目标形变和尺度变化，导致 RT-MDNET 精确率只有 80.1%，成功率只有 51.6%，本文算法适应了全景数据特性，通过采用 LSTM 算法减少尺度变化对目标跟踪网络产生的影响，降低了跟踪难度，从而提升了算法跟踪性能. 最终，本文算法精确率为 86.9%，成功率为 79.9%，速度也优于 ADNet 与 MDNet.

表 2 各算法在不同指标下性能对比

Tab. 2 Performance comparison of algorithms with different indicators

指标	阈值为 20 像素时的精确率	阈值为 0.5 时的成功率	平均中心位置误差	FPS
MDNet	0.744	0.562	50.22	1
ADNet	0.316	0.673	293.47	4
Siamese RPN	0.783	0.731	88.31	75
RT-MDNet	0.801	0.516	16.89	8
本文算法	**0.869**	**0.799**	**13.35**	7

综上所述,RT-MDNet 与 MDNet 都达到了很好的精度,但在应对尺度问题时难以适应变化程度. ADNet 在应对尺度变化的问题强于前者,但还是达不到对于全景数据的需求. Siamese RPN 较好地应对了尺度变化的问题,但相关滤波方法容易受到相似特征背景的影响导致精确率较低. 通过以上对比试验可以得出,经过 LSTM 网络的本文算法在主观标准和客观标准上均有很大的提升,在应对不同光照条件,不同目标时可以较好地应对目标的尺度变化和遮挡,并提高了在全景图像上的准确率和重叠率,跟踪效果明显提升.

3 结 论

为了解决基于全景数据集的目标跟踪的问题,本文提出了一种基于 RT-MDNet 和 LSTM 网络的全景图像跟踪算法,采用卷积神经网络提取特征,并利用 RoIAlign 方法来减少卷积过程中对特征区域的损耗,增强网络的鲁棒性;使用区分多视频序列间目标的损失函数,使网络可以更好的区分相似目标加强网络的适用性;设计 LSTM 网络自适应地选取边界框的尺度,针对数据集改进网络结构,以应对全景数据中出现的目标形变和尺度变化问题,最终输出目标位置信息.

实验结果表明,本文算法具有较高的跟踪精度,能够在目标扭曲、旋转剧烈、目标运动快、背景相似干扰等多种挑战下长期稳定地跟踪目标,在保持了精度的同时对全景数据的 IOU 得分实现了有效的提高. 但是由于全景图像分辨率较大的原因,伴随着运算量大的问题,导致算法速度受到限制,目前还难以满足实时的需求. 进一步裁剪网络、优化算法、实时处理将会是以后的重点研究方向.

参考文献

[1] 卢湖川, 李佩霞, 王栋. 目标跟踪算法综述[J]. 模式识别与人工智能, 2018, 31(1): 61
LU Huchuan, LI Peixia, WANG Dong. Overview of target tracking algorithms [J]. Pattern Recognition and Artificial Intelligence, 2018, 31(1): 61. DOI: 10.16451/j.cnki.issn1003 - 6059.201801006

[2] CAI Zhaowei, WEN Longyin, LEI Zhen, et al. Robust deformable and occluded object tracking with dynamic graph [J]. IEEE Transactions on Image Processing, 2014, 23(12): 5497. DOI:10.1109/TIP.2014.2364919

[3] NAM H, HAN B. Learning multi-domain Convolutional Neural Networks for visual tracking [C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV: IEEE, 2016: 4293. DOI:10.1109/CVPR.2016.465

[4] YUN S, CHOI J, YOO Y, et al. Action-decision networks for visual tracking with deep reinforcement learning [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017: 1349. DOI:10.1109/CVPR.2017.148

[5] 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, UT: IEEE, 2018:8971. DOI:10.1109/CVPR.2018.00935

[6] JUNG I, SON J, BAEK M, et al. Real-Time MDNet [M]// FERRARI V, HEBERT M, SMINCHISESCU C, et al. Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, vol 11208. Cham: Springer, 2018:89. DOI:10.1007/978-3-030-01225-0_6

[7] ZHOU Yuan, ZHOU Zhong, CHEN Ke, et al. Persistent object tracking in road panoramic videos[M]//LIN Weisi, XU Dong, HO A, et al. Pacific-Rim Conference on Multimedia. Berlin: Springer, 2012: 359. DOI:10.1007/978-3-642-34778-8_33

[8] REN Shaoqing, HE Kaiming, 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, 2015, 39(6): 1137. DOI:10.1109/TPAMI.2016.2577031

[9] HE Kaiming, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(1): 386. DOI:10.1109/TPAMI.2018.2844175

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735. DOI:10.1162/neco.1997.9.8.1735

[11] SHU Xiangbo, TANG Jinhui, QI Guojun, et al. Concurrence-aware long short-term sub-memories for person-person action recognition [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). San Francisco: IEEE, 2017: 2176. DOI:10.1109/CVPRW.2017.270

[12] WU Yi, LIM J W, YANG M H. Online object tracking: A benchmark[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR: IEEE, 2013: 23. DOI:10.1109/CVPR.2013.312

(下转第 174 页)