文章编号：2095-2163(2022)12-0159-06 中图分类号：TP391 文献标志码：A

基于特征聚合的模型预测跟踪方法

张 乐，韩 华，王春媛，马才良，王婉君，汤辰玉
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：为了提升基于模型预测的目标跟踪算法在复杂场景中的跟踪表现，提出了基于特征聚合的方法来获得更加具有判别力的鲁棒特征图，然后将该特征图送入模型预测器中对目标进行在线预测，最终能在多种复杂场景下实现实时鲁棒的跟踪任务。该方法的具体设计流程为：改进特征提取网络，并对特征提取网络的最后几层进行多层特征聚合操作。实验表明：所提出的算法在 VOT2018 数据集的 *EAO*（Expect Average Overlap）指标上比基线算法高了 4.88%；在 UAV123 数据集的成功率（*Success rate*）和精确率（*Precision rate*）指标上比基线算法分别提高了 4.5%和 4.4%。

关键词：特征聚合；模型预测；目标跟踪

Model prediction and tracking method based on feature aggregation

ZHANG Le, HAN Hua, WANG Chunyuan, MA Cailiang, WANG Wanjun, TANG Chenyu
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 In order to improve the tracking performance of object tracking algorithm based on model prediction in complex scenes, a method based on data augmentation and feature fusion is proposed. The method can obtain a more discriminative robust feature map, and then the feature map is sent to model predictor to carry out online prediction, finally realizes real-time robust tracking tasks in a variety of complex scenes. The specific design of the method is: feature extraction network is improved, after that multi-layer feature aggregation operation is performed on the last two layers of feature extraction network. Experiments show that the proposed algorithm is 4.88% higher than baseline algorithm on the *EAO* (Expected Average Overlap) indicator of VOT2018 dataset; it is higher than baseline on the *Success rate* and *Precision rate* indicators of UAV123 dataset by 4.5% and 4.4%, respectively.

【Key words】 feature aggregation; model prediction; object tracking

0 引 言

视觉目标跟踪是计算机视觉领域一个重要的研究方向。现已广泛地应用在公共安防[1-5]、自动跟踪[6]等方面。目标跟踪旨在当给定视频序列的第一帧的目标边界框的情况下，利用跟踪算法在视频序列的后续帧中定位该目标的准确位置，并同样使用边界框在视频帧中进行目标的框定。尽管目标跟踪领域在多方面探讨中已经取得了可观进展，然而在一些类似于光照变化、遮挡、背景干扰等场景中也亟待更深入系统的研究。

近年来，在目标跟踪方面涌现出众多的研究成果。尤其是基于暹罗（Siamese）网络[7]的跟踪算法，凭借着平衡的跟踪准确性和速度获得了相关学者极大的关注。暹罗网络的思想是将目标跟踪任务视为一个相似性匹配问题。具体来说，基于相似性匹配的跟踪方法是以端到端的方式从大量的数据集中离线学习一个通用的相似性匹配函数，训练目标是使同一个物体的相似性最大，不同物体的相似性最小。

尽管基于暹罗网络的跟踪算法已经取得不小进展，然而仍有改善和可提升空间。一方面，一些跟踪算法仅仅采用有限的数据增强策略，这对于训练一个鲁棒性的跟踪器是不够的。因此，有必要释放训练数据的潜力来训练跟踪算法。另一方面，基于暹罗网络的一般跟踪方法仅仅使用特征提取网络的最后一层的输出作为最终提取的特征图。这会导致跟踪器无法拥有比较强的判别能力。

为了解决上述问题，本文提出了一种特征聚合的模型预测目标跟踪方法。在模型层面，本文提出的多层特征聚合策略可以获得更加高质量的特征图。

基金项目：国家自然科学基金（61305014）；上海市自然科学基金（22ZR1426200）。

作者简介：张 乐（2000-），男，本科生，主要研究方向：目标识别与跟踪、图像处理、计算机视觉；韩 华（1983-），女，博士，教授，主要研究方向：目标识别与跟踪、行人重识别、智能计算等；王春媛（1983-），女，博士，副教授，主要研究方向：多源信息协同处理、模式识别、机器学习等。

通讯作者：韩 华 Email：2070967@mail.dhu.edu.cn

收稿日期：2022-05-08

1 本文算法研究

1.1 跟踪系统框架

本文的跟踪系统框架如图 1 所示。使用本文改进的特征提取网络进行特征的提取,并对特征提取网络的最后 2 个网络层进行特征聚合操作,以获得更加具有判别力的特征图。随后这些特征图进入模型预测模块中进行目标的在线更新操作。再将更新得到的模板作为一个卷积核与测试集的特征图进行卷积操作。最终,模型输出待跟踪目标的具体位置信息。

1.2 改进的特征提取网络

为了提高定位的准确性、降低计算量以及完成后续的多层特征聚合操作,本文对原始的特征提取网络 ResNet-50[8-9] 进行了如下的改进:

(1) 由于卷积操作中较大的步幅会降低定位的准确性,因此将特征提取网络中的第三和第四个卷积层的步幅大小从 2 设为 1。

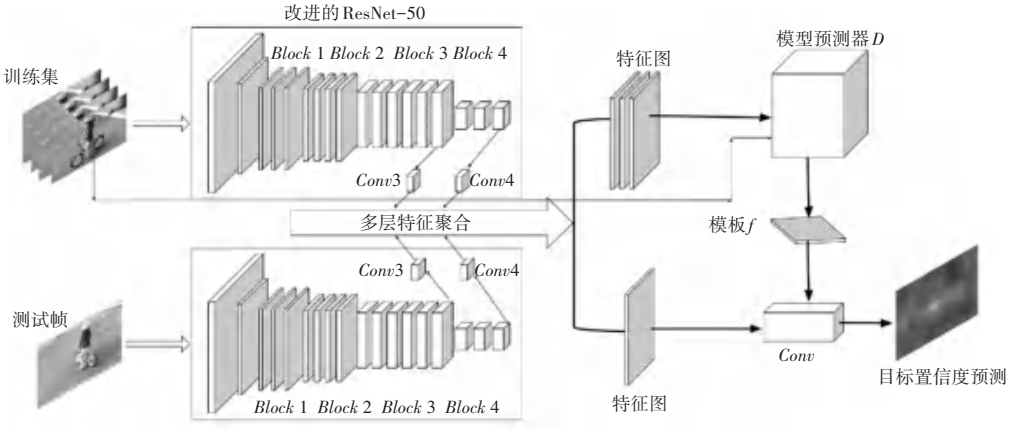


图 1 本文的跟踪系统框架

Fig. 1 The tracking system framework of this paper

(2) 由于特征聚合操作的网络层需要相同的通道数,故将第四个卷积层的通道数从 2 048 变为1 024。

(3)在第三和第四层的后面分别加上一个卷积核大小为 1×1 的卷积层来分别提取 2 个层的特征,并命名为 Conv3 和 Conv4。

改进的 ResNet-50 层级结构见表 1。

表 1 改进的 ResNet-50 层级结构

Tab. 1 Improved ResNet-50 hierarchical structure

Layer Name	Original layer structure	Modified layer structure
*Conv*1	7 × 7,64,2,1	7 × 7,64,2,1
Max pool	3 × 3,2,1	3 × 3,2,1
*Block*1	[1 × 1,64; 3 × 3,64; 1 × 1,256] × 3,2,1	[1 × 1,64; 3 × 3,64; 1 × 1,256] × 3,2,1
*Block*2	[1 × 1,128; 3 × 3,128; 1 × 1,512] × 4,2,1	[1 × 1,128; 3 × 3,128; 1 × 1,512] × 4,2,1
*Block*3	[1 × 1,256; 3 × 3,256; 1 × 1,1 024] × 6,2,1	[1 × 1,256; 3 × 3,256; 1 × 1,1 024] × 6,1,1
*Conv*3	Null	1 × 1,1 024,1,0
*Block*4	[1 × 1,512; 3 × 3,512; 1 × 1,2 048] × 3,2,1	[1 × 1,512; 3 × 3,512; 1 × 1,2 048] × 3,1,1
*Conv*4	Null	1 × 1,1 024,1,0

1.3 特征聚合策略

在目标跟踪领域,许多研究已经证明浅层的特征图包含目标更多的位置信息,深层的特征图包含目标更多的语义信息。这些语义信息对目标外观差异有着较好的不变性。

在目前研究发展基础上,本文提出了一个多层特征聚合策略,该策略将特征提取网络的最后 2 个特征提取层进行聚合来获得更加具有判别力的特征图。

本文提出的多层特征聚合框架图如图 2 所示。由图 2 可看到, *Conv*3 和 *Conv*4 分别用来提取卷积块 3 (*Block*3) 和卷积块 4 (*Block*4) 的特征。因此,一共可以获得 2 张特征图。

为了描述每张特征图的波动水平和感兴趣目标的置信度,本文使用公式(1)来计算每一个特征图的平均峰相关能量 (average peak-to-correlation-energy):

$$A_{PCE}=\frac{|\hat{V}-\check{V}|^2}{E\left(\sum_{w,h}(V_{w,h}-\check{V})^2\right)} \tag{1}$$

其中, $\hat{V}$ 是该特征图中的最大值; $\check{V}$ 是该特征图的最小值; $V_{w,h}$ 是矩阵 $\boldsymbol{V}$ 中第 w 行第 h 列对应的值; E 是算术平均算子。

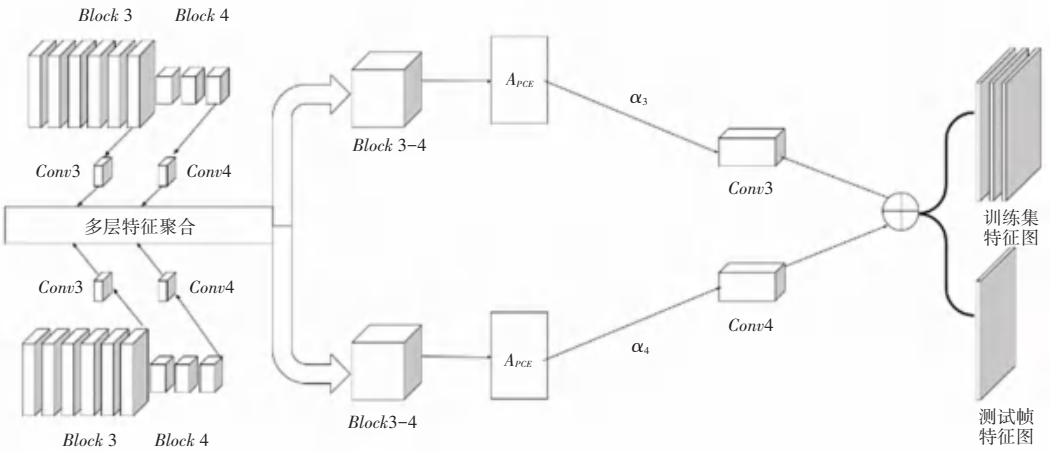


图 2　多层特征聚合框架图

Fig. 2　Multi-layer feature aggregation framework diagram

当计算 A_{PCE} 值后，每张特征图的权重可以由式(2)计算求得：

$$\alpha_i = \frac{A^i_{PCE}}{\sum_{i=3}^{4} A^i_{PCE}} \tag{2}$$

当计算 α_i 后，使用式(3)进行特征图的聚合：

$$F_{new} = \sum_{i=3}^{4} \alpha_i * \Omega_i \tag{3}$$

其中，Ω_i 为 $Conv(i)$ 输出的特征图。

2　实验与分析

2.1　实验参数设定

本文对算法的训练和评估参数进行了设置，具体参数如下。

(1)训练方案：使用 GOT10k[10] 和 LaSOT[11] 数据集的训练集部分作为数据集，并从这 2 个数据集中采样20 000个视频序列作为训练数据集。训练阶段的初始学习率为 0.001。优化器 ADAM 每 15 个世代（*epoch*）衰减 0.2。动量设置为 0.9，一共训练 50 个世代，整个训练的时长大约为 24 h。

(2)评估设计：本算法使用 VOT2018[12] 和 UAV123[13] 作为评估数据集并使用 PySOT 作为评估平台。首先生成本算法的.txt 格式跟踪结果，随后通过 PySOT 平台对不同的评估数据集进行评估，最终生成本文算法与不同跟踪算法的比较结果。

2.2　算法结果分析

为了量化所提出算法的跟踪表现，本文分别在 VOT2018 以及 UAV123 评估数据集上进行评测，并与其他具有竞争力的跟踪算法进行对比分析。

2.2.1　VOT2018 评估分析

VOT2018 由 60 个包含不同属性的 RGB 视频序列组成。与大多数研究者相似，本文使用 VOT 中的准确度（*A*）、鲁棒性（*R*）和平均期望均值（*EAO*）来评估不同的跟踪算法。*EAO* 作为一个跟踪算法最终的评估指标。通常 *EAO* 值越大，该跟踪算法的性能越好。表 2 为本文算法与 4 个具有竞争力的跟踪算法的对比结果。

表 2　VOT2018 上不同跟踪算法的比较

Tab. 2　Comparison of different tracking algorithms on VOT2018

	The proposed	DiMP	ATOM	SiamRPN++	DaSiamRPN
A	0.593	0.590	0.590	**0.600**	0.586
R	**0.154**	0.164	0.204	0.234	0.276
EAO	**0.430**	0.410	0.401	0.414	0.383

由表 2 分析可知，本文算法在对比的 4 个跟踪算法上表现居于第一。其仅仅在准确率上比 SiamRPN++算法低了 1.17%，但在鲁棒性和 *EAO* 指标上均优于对比的其他跟踪算法。而且本文的算法在 *EAO* 指标上比第二名 SiamRPN++算法高了 3.86%，比基线算法 DiMP（本文使用 LaSOT 和 GOT10k 的训练集训练 DiMP 算法得出的结果）高了 4.88%。这些结果充分证明了本算法的优势。

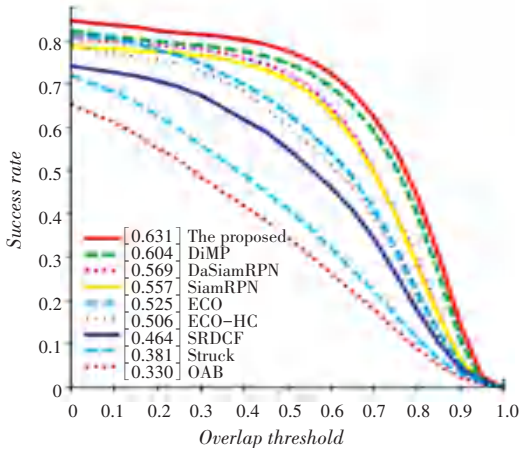
2.2.2　UAV123 评估分析

UAV123 数据集包含 123 个由低空无人机采集的视频序列。根据 UAV123 的评估标准，本文采用成功图（success plot）和精确图（precision plot）来对不同的算法进行比较。图 3 为不同跟踪算法在

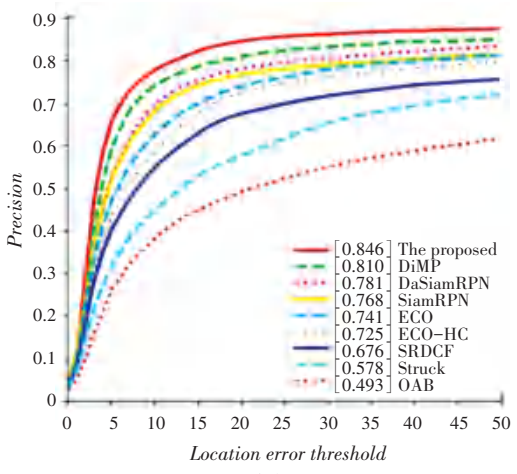
UAV123 上的成功率对比图和精确度对比图。图 4 为不同跟踪算法在 UAV123 数据集上 12 个不同跟踪属性的对比结果图。

由图 3 分析可知,本文所提出的算法在成功率和精确率方面均取得第一的位置。在成功率方面,DiMP 为 0.604,本文的算法成功率为 0.631,超过了第二名 DiMP 算法 4.5%。在精确率方面,本文的算法为0.846,超过了第二名 DiMP 算法 4.4%。这些结果充分说明了本文算法具有优秀的跟踪性能。

由图 4 可看到,本文的算法在 UAV123 所有 8 个跟踪属性上的表现均高于基线算法,且有 11 个跟踪属性都取得了第一的位置。这些结果说明了本文所提出的数据增强策略和特征聚合策略的有效性。



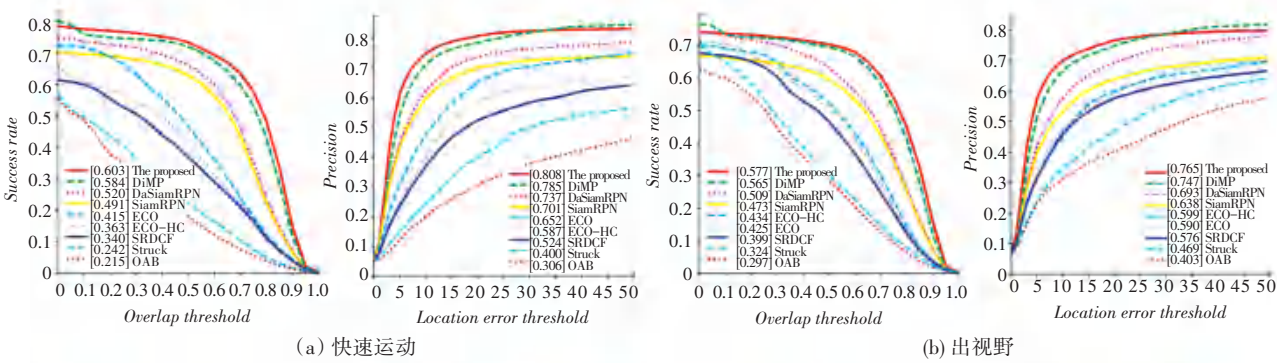
(a) 成功率



(b) 精度图

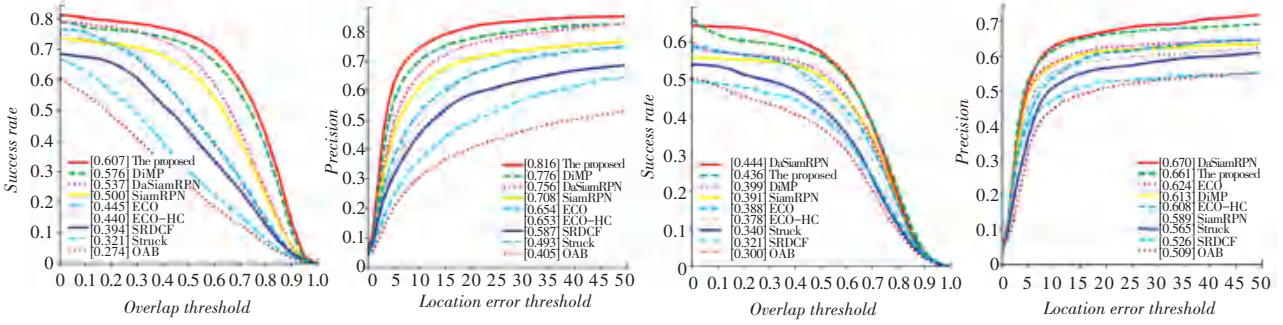
图 3 UAV123 上不同跟踪算法的比较

Fig. 3 Comparison of different tracking algorithms on UAV123



(a) 快速运动

(b) 出视野



(c) 长宽比变化

(d) 背景干扰

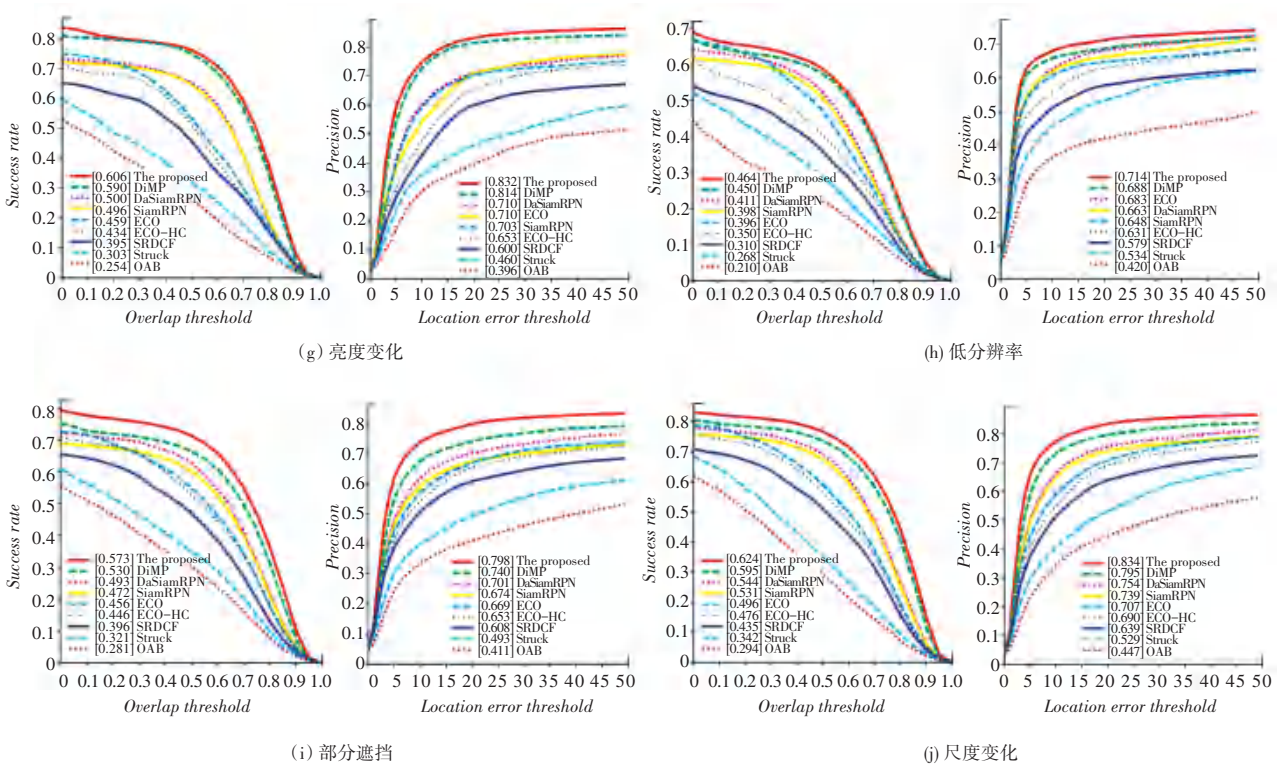


图 4　UAV123 不同属性的跟踪结果图

Fig. 4　Graph of tracking results for different attributes on UAV123

3　结束语

为了获得更加鲁棒性的特征图、从而在模型预测器中进行具有判别力的跟踪表现研究，本文分别从数据和模型两个方面进行改进。在数据方面，新引入了颜色抖动以及自定义了运动模糊数据增强方式；在模型方面，首先对特征提取网络 ResNet-50 进行了改进，然后在 ResNet-50 的最后 2 个特征提取层进行了特征聚合操作。最终训练的跟踪模型分别在 VOT2018 和 UAV123 数据集中进行了评估。在 VOT2018 上，本文的算法取得了第一的位置，并在 *EAO* 指标上比第二名算法高出了 3.86%，比基线算法 DiMP 高出了 4.88%。在 UAV123 上，本文的算法同样为最好的水平，同时在准确度和精确度上比第二名算法分别提高了 4.5%，4.4%。这些结果充分说明了本文所提出算法在跟踪方面有着更好的表现。

参考文献

[1] 唐佳敏，韩华，黄丽，等. 无监督行人重识别的判别性特征研究[J]. 智能计算机与应用，2021，11(08)：146-150.

[2] HAN Hua，MA Wenjin，ZHOU Mengchu，et al. A novel semi-supervised learning approach to pedestrian re-identification[J]. IEEE Internet of Things Journal，2021，8(4)：3042-3052.

[3] HAN Hua，ZHOU Mengchu，SHANG Xiwu，et al. KISS+ for rapid and accurate pedestrian re-identification[J]. IEEE Transactions on Intelligent Transportation Systems，2021，22(1)：394-403.

[4] HAN Hua，ZHOU Mengchu，ZHANG Yujin. Can virtual samples solve small sample size problem of KISSME in pedestrian re-identification of smart transportation?[J]. IEEE Transactions on Intelligent Transportation Systems，2020，21(9)：3766-3776.

[5] ZHANG Luyao，HAN Hua，ZHOU Mengchu，et al. An improved discriminative model prediction approach to real-time tracking of objects with camera as sensors[J]. IEEE Sensors Journal，2021，21(15)：17308-17317.

[6] 左国才，苏秀芝，陈明丽，等. 基于深度学习抗遮挡的多目标跟踪研究[J]. 智能计算机与应用，2020，10(07)：239-242.

[7] LI B，WU W，WANG Q，et al. Siamrpn++：Evolution of siamese visual tracking with very deep networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE，2019：4282-4291.

[8] BHAT G，DANELLJAN M，GOOL L V，et al. Learning discriminative model prediction for tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul，South Kerean：IEEE，2019：6182-6191.

[9] HE Kaiming，ZHANG Xiangyu，REN Shaoqing，et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas，NV，USA：IEEE，2016：770-778.

[10] 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.

（下转第 168 页）