文章编号：2095-2163(2022)12-0082-07 中图分类号：TP391 文献标志码：A

复杂场景下基于改进 DiMP 算法的精确目标跟踪

忻 瑶，韩 华，王春媛，熊雨滋，许莹莹
（上海工程技术大学 电子电气工程学院，上海 201620）

摘 要：针对 DiMP 目标跟踪算法在自然场景下遇到遮挡及背景干扰导致跟踪表现不佳的问题，提出了改进的 DiMP 精确目标跟踪算法。在图像预处理阶段创新性地设计了一个任意灰度块替换策略来丰富样本的信息；将特征提取网络 ResNet-50 提取的目标各阶段的特征图输入到设计的多尺度融合模块中进行正向和反向的充分融合，得到包含更多位置信息和语义信息的特征图；随后特征图输入到模板预测模块中进行在线更新操作，进而得到判别力更强的目标模板。实验表明：该算法在 UAV123 数据集的遮挡和背景干扰测试中的成功率和精确率分别提高 8%、4.15%和 9%、6.30%；同时，在 VOT2018 的 *EAO* 指标上提高 1.36%，在 UAV123 的成功率和精确率指标上分别提高 3.89%和 3.06%。说明改进的 DiMP 算法在对遮挡与背景干扰问题上优势明显，进而提升了算法的整体表现。

关键词：替换策略；多尺度融合模块；DiMP；目标跟踪

Accurate object tracking based on improved DiMP algorithm in complex scenes

XIN Yao, HAN Hua, WANG Chunyuan, XIONG Yuzi, XU Yingying
(School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

【Abstract】 Aiming at the problem that DiMP object tracking algorithm encounters occlusion and background interference in natural scenes, which leads to poor tracking performance, an improved DiMP accurate object tracking algorithm is proposed. In image preprocessing stage, an arbitrary gray-scale block replacement strategy is innovatively designed to enrich the information of samples; the feature maps of each stage of objects extracted by ResNet-50 are input into designed multi-scale fusion module for full fusion of forward and reverse directions, and feature maps containing more location information and semantic information are obtained; then feature map is input into template prediction module for online update operation, and target template with stronger discriminative power is obtained. Experiments show that success rate and precision rate in the occlusion and background interference tests of UAV123 dataset are increased by 8%, 4.15% and 9%, 6.30%, respectively; meanwhile, *EAO* indicator of VOT2018 is increased by 1.36%, success rate and precision rate of UAV123 are increased by 3.89% and 3.06%, respectively. The simulation shows that improved DiMP has a higher advantage in occlusion and background interference problems, thereby improving the overall performance of the algorithm.

【Key words】 replacement strategy; multi-scale fusion module; DiMP; object tracking

0 引 言

视觉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同样以边界框的形式自动定位该目标的准确位置。作为计算机视觉领域一个重要的研究方向，现已广泛地应用在视频公共安防[1-5]、自动驾驶[6]、无人机[7]、机器人[8]等方面。

近年来，很多学者已经在深度学习目标跟踪方面取得了可观的研究成果。2016 年，Bertinetto 等人[9]以端到端的方式成功训练了第一个全卷积暹罗网络并命名为 SiamFC，该算法不仅推理速度可达实时，同时表现出优良的跟踪性能。2018 年，Li 等人[10]在 SiamFC 的基础上将目标检测中的区域建议网络(region proposal network, RPN)[11]引入到目标跟踪领域，RPN 模块可以使跟踪器回归位置、形状，省掉多尺度测试环节，所提出的 SiamRPN 算法进一步提高了跟踪速度(160 FPS)，并且拥有更高的跟踪准确度和精确度。2019 年，Li 等人[12]将特征提取网络替换成层数更深、拟合能力更强的 ResNet[13]网络，成功训练了以 ResNet 为驱动的 SiamRPN++。然而，这些 Siamese 类跟踪算法仅仅利用了目标的

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：忻 瑶(2000-)，女，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：2022-01-17

外观信息，未将背景考虑进去，并且未对目标模板进行在线更新。当遇到复杂背景或目标发生严重畸变的情况下，Siamese 类算法很容易发生跟踪漂移的情况。2019 年，Bhat 等人[14]通过联合目标的外观和背景信息并通过在线更新的方式获得具有判别力的目标模板，不仅实现了实时的跟踪速度，而且所提出的 DiMP 算法在多个评估数据集上均取得第一的位置。

尽管这些跟踪算法已经取得了很大进展，不断刷新跟踪表现，然而仍然有不少缺陷。一方面，DiMP 算法仅仅采取通用的数据增强策略，比如任意裁剪、旋转等，跟踪算法只能学到有限的数据信息。因此，有必要做更适合目标跟踪的数据增强来释放数据的潜力。另一方面，DiMP 算法仅仅使用特征提取网络的最后一层的输出作为目标的特征图，未能使特征图包含充分的语义和位置信息。

为了提高 DiMP 算法在面对目标遮挡、背景干扰场景下的跟踪表现，本文在数据预处理阶段设计了一个高效的任意灰度块替换策略，在特征提取网络后面添加了一个多尺度融合模块。具体贡献如下：

(1)设计了一个任意灰度块替换策略使数据样本模拟真实场景中的目标遮挡、光线变化的情况，增加样本的多样性，降低遮挡、光线变化导致的模型过拟合的风险。

(2)设计了一个多尺度特征融合模块，该模块对特征提取网络提取的不同阶段的特征图进行正向和反向的多尺度融合，得到语义信息和位置信息更强的目标特征图。

(3)在主流的评估数据集上进行评测分析，验证了改进的 DiMP 算法在遮挡、背景干扰场景下有更好的跟踪表现。

1　本文算法

1.1　跟踪系统框架

为了提升 DiMP 算法在遮挡、背景干扰场景下的跟踪表现，本文探索并改进了 DiMP 算法[15]。改进的 DiMP 算法由 5 部分组成，如图 1 所示。图 1 中，第 1 部分是输入端，由训练分支和测试分支组成。输入到训练分支的图片为经过本文任意灰度块替换策略后的训练样本；第 2 部分是特征提取网络 ResNet-50，用来提取跟踪目标各个阶段的多尺度特征图；第 3 部分是本文提出的多尺度融合模块，该模块由上采样子模块和下采样子模块组成，用来对特征提取网络输出的各个阶段特征图进行正向和反向的多尺度特征融合，得到语义信息和位置信息更加充分的特征图；第 4 部分为模型预测模块，目标特征图和对应的边界框真值同时输入到该模块中进行不断在线更新，得到目标模板；第 5 部分为互相关模块，目标模板作为卷积核与经过测试分支得到的特征图进行互相关操作，得到目标的置信度预测。

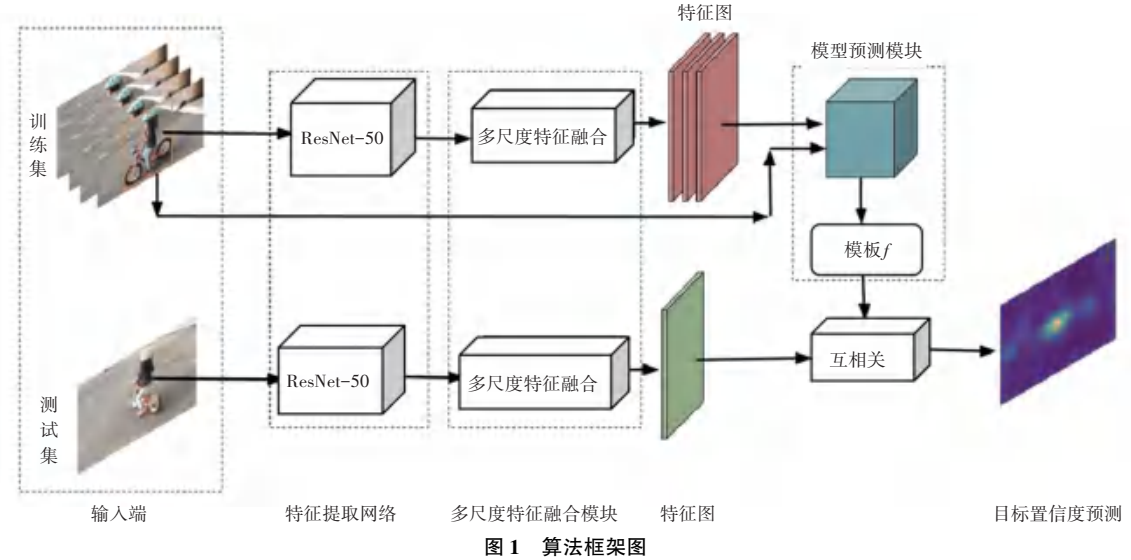


图 1　算法框架图

Fig. 1　The pipeline of the algorithm

1.2　任意灰度块替换策略

在实际的跟踪中，目标可能会出现部分遮挡、光线变化等影响跟踪的情况。因此，本文创新性地设计了一个任意灰度块替换策略。该策略随机选择图

像中的一个矩形区域,并用相应灰度图像中相同的矩形区域进行像素替换,从而生成灰度块替换后的训练样本。

在该方法中,进行任意灰度块替换的概率为 p,随机生成的矩形区域与图像的面积之比的最小值和最大值分别为 $S_{\min}$ 和 $S_{\max}$,矩形区域的面积 S_r 的取值范围为 $Rand(S_{\min},S_{\max}) \times S$。$\gamma$ 为确定矩形形状的系数,取值范围为 $[\gamma_1,\gamma_2]$,x_r 和 y_r 为矩形左上角的位置坐标,当矩形的坐标超过图像范围,需重新确定矩形的位置坐标。

该策略可以很好模拟自然场景中由于图像分辨率低或者光线变化导致的颜色变化问题,同时模拟目标遇到的部分遮挡问题。并且,该策略可以在保留图片结构信息的基础上增加样本多样性。设计的任意灰度块替换策略效果如图 2 所示。图 3 为任意灰度块替换策略在网络中的使用图。

图 2 任意灰度块替换策略效果图

Fig. 2 Arbitrary gray block replacement strategy renderings

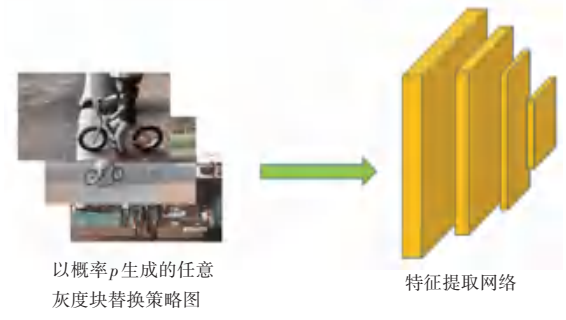


图 3 任意灰度块替换策略在网络中的使用图

Fig. 3 Diagram of arbitrary gray-scale block replacement strategy in networks

1.3 多尺度特征融合模块

为了获得融合目标语义信息与位置信息的特征图,本文在特征提取网络后设计了一个多尺度特征融合模块。该多尺度特征融合模块由上采样子模块和下采样子模块组成。

研究中给出的多尺度特征融合模块如图 4 所示。特征提取网络对预处理后的训练样本进行特征提取,生成各阶段的目标特征图,即 $\{C_2, C_3, C_4, C_5\}$;上采样子模块通过上采样和正向连接操作将特征提取网络的特征图进行自顶向下的多尺度融合,C_5 经 $1 \times 1 \times 256$ 卷积操作得到 T_5,随后 T_5 经过二倍上采样的结果与相邻的下层特征图 C_4 经过 $1 \times 1 \times 256$ 卷积操作得到的结果进行张量相加得到 T_4。T_3 和 T_2 的获取流程同 T_4,最终上采样子模块得到 $\{T_2, T_3, T_4, T_5\}$,其中 T_2、T_3 和 T_4 均融合了本层和更高层的信息。随后,下采样子模块通过下采样和反向链接操作将 $\{T_2, T_3, T_4, T_5\}$ 进行自底向上的多尺度融合,T_2 经过 $1 \times 1 \times 256$ 卷积得到 D_2,D_2 经过两倍下采样与相邻的上层特征图 T_3 进行张量相加得到 D_3,D_4 和 D_5 的获取流程同 D_3,最终下采样得到语义信息和位置信息更强的 $\{D_2, D_3, D_4, D_5\}$,其中 D_5 充分融合了多尺度特征图中的语义信息和位置信息,可作为多尺度特征融合模块最终的输出特征图。

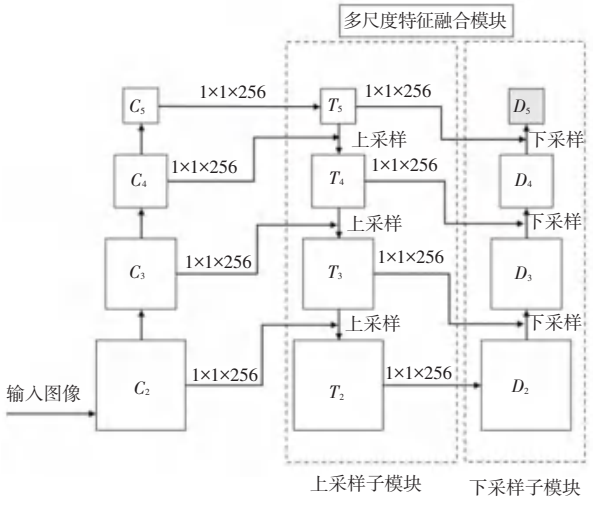


图 4 多尺度特征融合模块

Fig. 4 Multi-scale feature fusion module

2 实验与分析

2.1 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)训练方面:本文使用 GOT10k[16] 和 LaSOT[17] 共 2 个数据集,并从这 2 个数据集中随机采样 20 000 个视频序列作为训练数据集。采用 PyTorch 深度学习框架,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,p 的值设为 0.4,一共训练 50 个世代,通过 RTX 1080ti 显卡进行训练,整个训练的时长大约为 24 h。

（2）评估方面：本算法使用 VOT2018[18] 和 UAV123[19] 作为评估数据集，并使用商汤开源的 PySOT 平台进行评估。这里先由不同的跟踪器生成 .txt格式跟踪边界框坐标，随后通过 PySOT 平台对不同的跟踪算法进行评估，最终生成本文改进 DiMP 算法与多个不同跟踪算法的比较结果。

2.2　算法结果分析

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（Accuracy，A）、鲁棒性（Robustness，R）和期望平均覆盖率（Expected Average Overlap，EAO）来评估不同的跟踪算法。其中，准确度的定义为预测框与真实框之间的交并比（Intersection-over-Union，IoU）。鲁棒性定义为跟踪算法在一个视频序列中跟踪失败的次数，单帧准确度的值低于设定的阈值即视为失败。期望平均覆盖率作为评估一个跟踪算法的最终指标，按照该指标的大小进行排名。通常期望平均覆盖率值越大，表明该跟踪算法的性能越好。研究推得的数学定义式可表示为：

$$\varphi_{N_s}=\frac{1}{N_s}\sum_{i=1:N_s}\varphi_i \tag{1}$$

其中，N_s 为一个视频总帧数，φ_i 为第 i 帧的准确度。

表 1 为本算法与 4 个其他具有竞争力的算法的对比结果。通过表 1 可以看出，本文改进的 DiMP 算法在性能表现上要优于做基准对比的 4 个跟踪算法。在准确率指标上，比第二名算法 SiamRPN++算法高了 1.17%，比 DiMP 算法高了 1.68%。在鲁棒性指标上，比 DiMP 算法高了 2.61%。而且改进的 DiMP 算法在 EAO 指标上比第二名 DiMP 算法高了 1.36%，比 SiamRPN++算法高了 7.73%。这些结果充分证明了改进的 DiMP 算法有着更好的跟踪表现。

表 1　VOT2018 上不同跟踪算法的比较

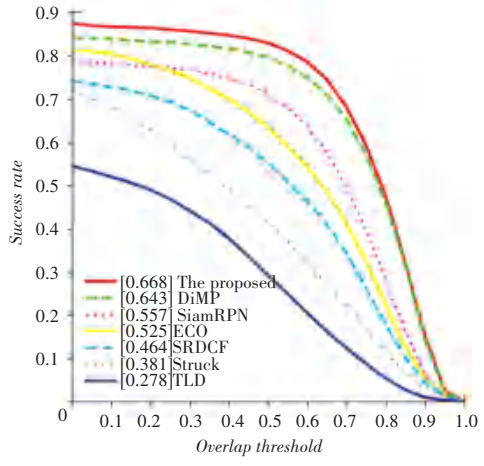
Tab. 1　Comparison of different tracking algorithms on VOT2018

算法	A	R	EAO
The proposed	**0.607**	**0.149**	**0.446**
DiMP	0.597	0.153	0.440
ATOM [20]	0.590	0.204	0.401
SiamRPN++[12]	0.600	0.234	0.414
DaSiamRPN [21]	0.586	0.276	0.383

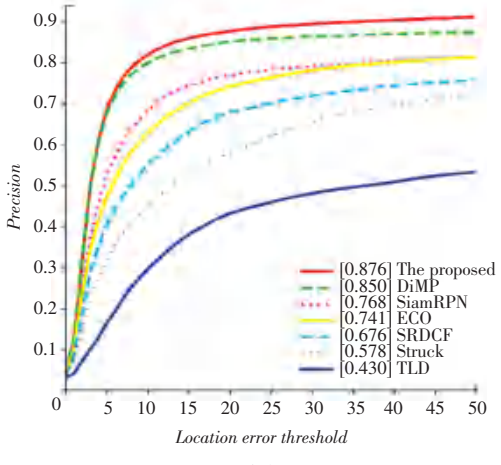
2.2.2　UAV123 评估分析

UAV123 数据集包含 123 个由低空无人机采集的视频序列。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 5 为不同跟踪算法在 UAV123 上的成功率对比图和精确度对比图。由图 5 可以看出，本文所提出的算法在成功率和精确率方面均为最佳。在成功率方面，本文算法的成功率为 0.668，超过了第二名 DiMP 算法 3.89%。在精确率方面，本文的算法为 0.876，超过了第二名 DiMP 算法 3.06%。这些结果充分说明了本文算法具有优秀的跟踪性能。



（a）成功图

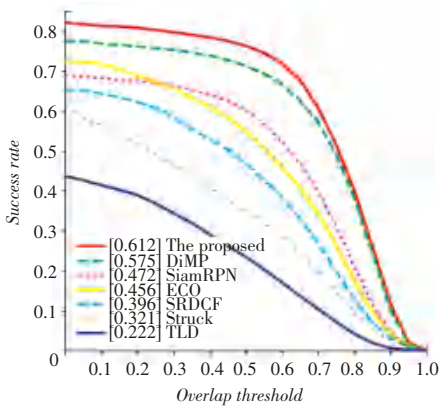


（b）精度图

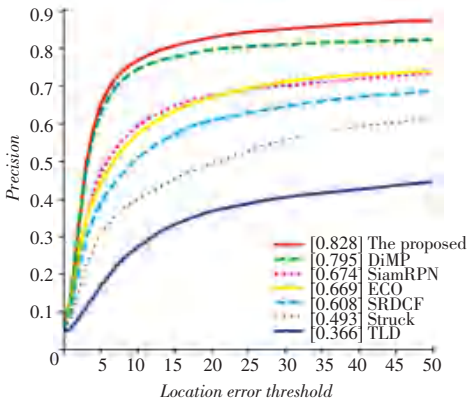
图 5　UAV123 上不同跟踪算法的比较

Fig. 5　Comparison of different tracking algorithms on UAV123

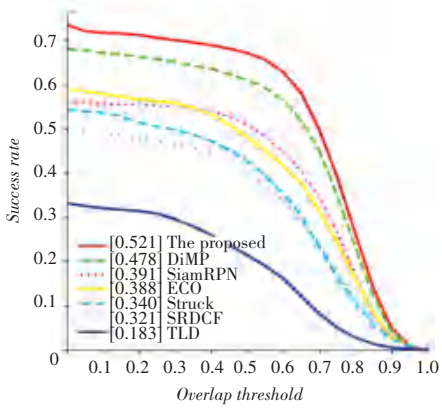
图 6 为不同跟踪算法在 UAV123 数据集的遮挡和背景干扰跟踪场景的对比结果图。由图 6 可以看出，本文改进的 DiMP 算法在遮挡场景中的成功率和精确率达到了 0.612 和 0.828，性能大幅度超过了原 DiMP 算法。同时，改进的 DiMP 算法在背景干扰场景中的成功率和精确率分别为 0.521 和 0.759，同样优于原 DiMP 算法结果。



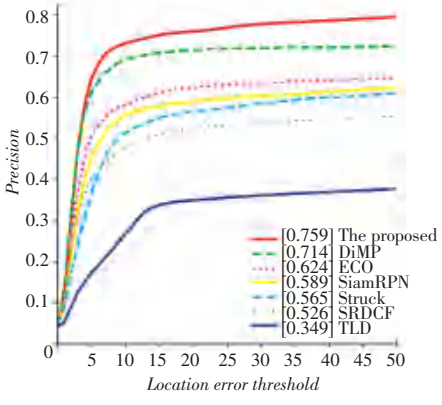
(a) 部分遮挡成功图



(b) 部分遮挡精度图



(c) 背景干扰成功图



(d) 背景干扰精度图

图 6 不同算法在遮挡、背景干扰场景下的表现

Fig. 6 The performance of different algorithms in occlusion and background interference scenes

2.2.3 实际场景跟踪分析

为了可视化本文改进的 DiMP 算法和基线算法在实际面对遮挡、背景干扰情况下的跟踪区别,本小节采集了一段包含遮挡和背景干扰的视频,并使用改进 DiMP 算法和原 DiMP 算法进行可视化分析,如图 7 所示。

图 7 实际的跟踪场景分析图

Fig. 7 Actual tracking scene analysis diagram

在图 7 中,红色框为改进 DiMP 的跟踪结果,黄色框为原始的 DiMP 算法跟踪结果。在第 5 帧目标基本无干扰的情况下,2 个算法的跟踪结果大体一致。当在第 138 和 270 帧时,目标遇到部分遮挡问题,可以看出,改进的 DiMP 算法可以很好地跟踪目标,而原始的 DiMP 算法的跟踪目标框已经出现了不准确的情况。另外,当目标在 251 帧出现严重背景干扰的情况下,DiMP 算法出现了跟踪漂移,而改进 DiMP 算法依然可以实现鲁棒性的跟踪。

3 消融实验分析

本文通过提出任意灰度块替换策略以及设计多尺度特征融合模块,使改进的 DiMP 算法在面对遮挡和背景干扰场景中有着更加鲁棒性的表现。下面通过消融实验分析所设计的策略和模块的影响,并在 VOT2018 和 UAV123 数据集上分别进行评估,结果见表 2。

在表 2 中,DiMP 表示原 DiMP 算法,DiMP+灰度块替换表示采用任意灰度块替换策略,DiMP+多尺度融合表示多尺度融合模块,改进 DiMP 算法表示采用任意灰度块替换策略和多尺度融合模块后的 DiMP 算法。S – 遮挡、P – 遮挡表示在遮挡和场景下

的成功率和精确率，S – 背景干扰、P – 背景干扰表示在背景干扰场景下的成功率和精确率。

表 2　消融实验分析

Tab. 2　Analysis of ablation experiments

	VOT2018	UAV123			
		S – 遮挡	P – 遮挡	S – 背景干扰	P – 背景干扰
DiMP	0.440	0.575	0.795	0.478	0.714
DiMP+灰度块替换	0.442	0.580	0.803	0.492	0.720
DiMP+多尺度融合	0.445	0.584	0.810	0.513	0.748
改进 DiMP	0.446	0.612	0.828	0.521	0.759

可以看出，任意灰度块替换策略和多尺度特征融合模块分别在 VOT2018 数据集上都有小幅的性能提升，在 UAV123 的遮挡和背景干扰场景下的成功率和精确率均有所提高。另外，相较于任意灰度块替换策略，多尺度特征融合模块对遮挡和背景干扰场景有着更大的贡献。这些结果说明了本文改进的 DiMP 算法在遮挡和背景干扰方面有着更好的跟踪精度。

4　结束语

本文针对 DiMP 算法在遮挡和背景干扰场景下表现不佳的问题，在数据预处理阶段设计了一个任意灰度块替换策略来应对光照变化和遮挡问题，以及在特征提取网络后设计了一个多尺度融合模块使各个阶段的特征图进行充分的融合。训练的跟踪模型在 VOT2018 和 UAV123 数据集上均取得总体表现第一的位置。并且在 UAV123 的遮挡和背景干扰场景下均优于其他跟踪算法，这些结果充分说明了本文改进的 DiMP 算法对目标遮挡和背景干扰场景有着更好的表现。

参考文献

[1] 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.

[2] 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.

[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] 张子龙，王永雄. 基于卡尔曼滤波的 SiamRPN 目标跟踪方法[J]. 智能计算机与应用, 2020, 10(03): 44-50.

[6] XIONG Zuobin, LI Wei, Han Qilong, et al. Privacy-preserving auto-driving: a GAN-based approach to protect vehicular camera data[C]//2019 IEEE International Conference on Data Mining (ICDM). Beijing, China: IEEE, 2019: 668-677.

[7] FAN Heng, WEN Longyin, DU Dawei, et al. VisDrone-SOT2020: The vision meets drone single object tracking challenge results[C]//European Conference on Computer Vision. Cham: Springer, 2020: 728-749.

[8] HEINRICH S, SPRINGSTÜBE P, KNÖPPLER T, et al. Continuous convolutional object tracking in developmental robot scenarios[J]. Neurocomputing, 2019, 342: 137-144.

[9] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking[C]//European Conference on Computer Vision. Cham: Springer, 2016: 850-865.

[10] 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, USA: IEEE, 2018: 8971-8980.

[11] REN Shaoqing, HE Kaiming, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. arXiv preprint arXiv: 1506.01497, 2015.

[12] LI Bo, WU Wei, WANG Qiang, 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, CA, USA: IEEE, 2019: 4282-4291.

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

[14] 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, Korea (South): IEEE, 2019: 6182-6191.

[15] 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.

[16] 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.

[17] FAN Heng, LIN Liting, YANG Fan, et al. LaSOT: A high-quality benchmark for large-scale single object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 5374-5383.

[18] KRISTAN M, LEONARDIS A, MATAS J, et al. The sixth visual object tracking vot2018 challenge results[M]//LEAL-TAIXÉ L, ROTH S. Computer Vision-ECCV 2018 Workshops. ECCV 2018. Lecture Notes in Computer Science( ). Cham: Springer, 2018, 11129: 3-53.

[19] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for uav tracking[C]//European Conference on Computer Vision. Cham: Springer, 2016: 445-461.

（下转第 92 页）