孪生网络框架下融合显著性和干扰在线学习的航拍目标跟踪算法

孙 锐 方林凤* 梁启丽 张旭东

(合肥工业大学计算机与信息学院 合肥 230009)

(工业安全与应急技术安徽省重点实验室 合肥 230009)

摘 要：针对一般跟踪算法不能很好地解决航拍视频下目标分辨率低、视场大、视角变化多等特殊难点，该文提出一种融合目标显著性和在线学习干扰因子的无人机(UAV)跟踪算法。通用模型预训练的深层特征无法有效地识别航拍目标，该文跟踪算法能根据反向传播梯度识别每个卷积滤波器的重要性来更好地选择目标显著性特征，以此凸显航拍目标特性。另外充分利用连续视频丰富的上下文信息，通过引导目标外观模型与当前帧尽可能相似地来在线学习动态目标的干扰因子，从而实现可靠的自适应匹配跟踪。实验证明：该算法在跟踪难点更多的UAV123数据集上跟踪成功率和准确率分别比孪生网络基准算法高5.3%和3.6%，同时速度达到平均28.7帧/s，基本满足航拍目标跟踪准确性和实时性需求。

关键词：目标跟踪；无人机航拍场景；孪生网络；目标显著性；在线学习干扰因子

中图分类号：TN911.73; TP391 **文献标识码**：A **文章编号**：1009-5896(2021)05-1414-10

DOI: 10.11999/JEIT200140

Siamese Network Combined Learning Saliency and Online Leaning Interference for Aerial Object Tracking Algorithm

SUN Rui FANG Linfeng LIANG Qili ZHANG Xudong

(*School of Computer and Information*, *Hefei University of Technology*, *Hefei* 230009, *China*)

(*Anhui Province Key Laboratory of Industry Safety and Emergency Technology*, *Hefei* 230009, *China*)

Abstract: In view of the fact that the general tracking algorithm can not solve the special problems such as low resolution, large field of view and many changes of view angle, a Unmanned Aerial Vehicle (UAV) tracking algorithm combining target saliency and online learning interference factor is proposed. The deep feature that the general model pre-trained can not effectively identify the aerial target, the tracking algorithm can better select the salient feature of each convolution filter according to the importance of the back propagation gradient, so as to highlight the aerial target feature. In addition, it makes full use of the rich context information of the continuous video, and learn the interference factor of the dynamic target online by guiding the target appearance model as similar as possible to the current frame, so as to achieve reliable adaptive matching tracking. It is proved that the tracking success rate and accuracy rate of the algorithm are 5.3% and 3.6% higher than that of the siamese network benchmark algorithm on the more difficult UAV123 dataset, respectively, and the speed reaches an average of 28.7 frames per second, which basically meet the aerial target tracking accuracy and real-time requirements.

Key words: Object tracking; Unmanned Aerial Vehicle (UAV) scene; Siamese network; Target saliency; Online leaning interference factor

收稿日期：2020-03-03；改回日期：2020-10-21；网络出版：2020-11-19

*通信作者： 方林凤 f_linf@163.com

基金项目：国家自然科学基金(61471154, 61876057)，安徽省重点研发计划-科技强警专项(202004d07020012)

Foundation Items: The National Natural Science Foundation of China (61471154, 61876057), The Key Research Plan of Anhui Province - Strengthening Police with Science and Technology (202004d07020012)

1　引言

随着无人机和计算机视觉的快速发展，基于无人机的智能目标跟踪系统在目标监控、军事反恐侦察等各个领域均有广泛应用[1,2]。航拍视频具有信息量大、背景复杂、视场不确定、跟踪目标小等特点，而现有目标跟踪算法没有完全针对这些特点进行设计和优化，所以在航拍视频中实现鲁棒且实时的跟踪仍然是一个巨大的挑战。

现有主流的目标跟踪算法都是基于深度学习的，它们主要分为两类[3]：第1类使用用于目标识别任务预先训练的深度模型来提取特征，将目标跟踪问题转化为分类问题。很多研究者在大型分类数据集上(比如ImageNet)直接训练网络，并采用预训练的卷积神经网络(Convolutional Neural Network, CNN)来提取目标特征，然后基于分类任务的观察模型得到跟踪结果。这种深度学习跟踪方法更多地强调设计有效的观测模型，虽然对各种观测模型，如相关滤波器[4]、回归器[5,6]和分类器[7]进行了广泛的研究，但对学习有差别的深层特征的关注却很少。通过分析发现，当使用经过预先训练的深层特征作为目标表示时，可能会出现许多问题。首先，由于分类任务更多地关注类之间的对象分类，针对分类任务预先训练的网络忽略了类内的差异。其次，即使目标对象出现在预先训练模型的训练集中，从最后的卷积层中提取的深层特征通常只保留较高层次的视觉信息，而这些信息对于精确定位或尺度估计来说并不那么有效。最后，深度目标跟踪算法[8]需要较高的计算负荷，因为预先训练模型的深层特征是高维的。为了缩小这一差距，利用与目标专门相关的深层特征进行航拍视频跟踪是非常重要的。

第2类成熟策略是基于匹配的跟踪策略，将候选样本与目标模板匹配，不需要在线更新。这种跟踪算法最显著的优点是其实时性[9]。最近，基于匹配的跟踪还可以使用深度模型来提高匹配的泛化能力[3,10,11]。通过学习一个通用的匹配函数，以保持实时响应能力。最近成功的模型有全卷积孪生网络跟踪算法[9](Fully-Convolutional Siamese networks, SiamFC)，虽然它不仅实现了不错的跟踪精度还满足了实时性，但是SiamFC[9]缺乏一个有效的在线更新模型去捕捉航拍场景下目标、背景或成像条件的时间变化。为了解决这一问题，文献[12]提出了使用新模板的线性交互来进行模型更新。文献[13]提出采用双模板的方式进行跟踪，使用改进的APECs更新策略进行模板更新，但只在准确性方面稍有提升。为了进一步提高SiamFC[9]的速度，文献[14]利用深度强化学习训练策略，在响应置信度足够高的情况下尽早停止卷积神经网络(CNN)的前馈计算，从而降低了简单帧的特征计算成本。这些方法往往具有较高的计算复杂度，且不能很好地克服航拍场景下目标跟踪的难点。本文模型充分利用视频序列丰富的上下文信息，可以有效地在线学习无人机下动态目标的外观干扰变化，实现可靠、有效、实时的自适应匹配跟踪。

针对以上深度目标跟踪算法的分析，本文提出了基于航拍目标显著性和在线自适应匹配的动态孪生网络跟踪算法。主要贡献如下：(1)针对用于分类任务预训练的网络很难学习有差别的深层特征进行航拍目标跟踪的问题，本文设计了一种新的特征通道回归损失函数来学习目标智能感知的深层特征，从而可以选择出最有效的卷积滤波器来生成航拍目标显著性特征，大大减少通道特征量来加速跟踪过程。(2)针对SiamFC[9]等算法无法在线更新目标模板从而导致目标发生显著变化时跟踪失败的问题，本文采用目标干扰因子在线学习机制来抑制航拍目标显著变化所带来的影响，从而实现有效可靠的自适应匹配跟踪。(3)通过在数据集上的大量实验证明，在无人机场景下，本文所提出的目标显著性特征的动态孪生网络跟踪算法具有比较有竞争力的性能，在UAV123数据集上跟踪成功率和准确率分别比孪生网络基准算法高5.3%和3.6%，同时速度达到平均28.7帧/s，基本满足航拍目标跟踪准确性和实时性需求。

2　本文跟踪算法

视觉目标跟踪就是在缺少目标先验知识的前提下，给定某视频序列初始帧的目标大小与位置，并在后续的视频帧中预测该目标的位置。基于模板匹配的孪生网络跟踪算法是一种监督学习算法，是通过共享权值来实现的。最经典的孪生网络跟踪算法SiamFC[9]的表达式为

$$\boldsymbol{f}(\boldsymbol{z},\boldsymbol{x})=\varphi(\boldsymbol{z})*\varphi(\boldsymbol{x})+\mathbf{b}\mathbf{I} \tag{1}$$

其中，$\boldsymbol{z}$代表的是模板图像，$\boldsymbol{x}$是待搜索图像，$\boldsymbol{f}(\boldsymbol{z},\boldsymbol{x})$是表示$\boldsymbol{z}$和$\boldsymbol{x}$中候选块之间的相似性的响应图；$\varphi(\cdot)$代表CNN深度特征；本文用$*$表示卷积操作，$\mathbf{b}\mathbf{I}$表示每个位置的取值。

图1展示的是本文跟踪算法的框架，包括预先训练的特征提取器、目标显著性感知特征模块、目标干扰因子在线学习模块和孪生网络匹配模型。因此，本文将原有的静态模型式(1)扩展到目标显著性感知的孪生网络动态匹配过程中

$$\boldsymbol{G}_t=\mathrm{corr}(\boldsymbol{S}_{t-1}*\boldsymbol{f}(\boldsymbol{z}),\chi'(\boldsymbol{x}_t)) \tag{2}$$

其中，$\boldsymbol{G}_t$是在第t帧图像中各个位置与模板图像的

响应值；*表示可以在频域快速求解的卷积。与式(1)不同的是。本文引入干扰因子$\boldsymbol{S}_{t-1}$来更新目标受到干扰的变化，通过引导第1帧的模板与当前帧的前一帧尽可能地相似来在线学习干扰因子$\boldsymbol{S}_{t-1}$。χ'提取的是待搜索图像的目标显著性感知特征，根据反向传播梯度判断每个卷积核用于描述物体特征的重要性来生成目标显著性特征。本文算法的可视化跟踪结果如图2所示，其中的颜色是将预测目标区域进行HSV颜色空间转换。

2.1 预训练网络的目标特征

一般视觉识别任务和特定目标跟踪在卷积神经网络提取的特征的有效程度是不同的。关于神经网络解释的几种方法证明了卷积滤波器对捕获类别级对象信息的重要性可以通过相应的梯度[15]来计算，由此本文构建了目标显著性感知特征模块，该模块具有专门针对视觉跟踪而设计的损失函数。给定具有输出特征空间的预先训练的CNN特征提取器，可以根据通道重要性生成目标显著性感知特征

$$\chi' = \rho(\boldsymbol{x}; \boldsymbol{M}) \tag{3}$$

其中，ρ为选择重要特征通道的函数，第i个特征通道的重要性可由式(4)计算得出

$$M_i = F_{\text{ap}}\left(\frac{\partial J}{\partial F_i}\right) \tag{4}$$

其中，F_{ap}表示全局平均池化函数，J是设计的损失函数，F_i为第i个滤波器的输出特征。对于航拍视频的目标跟踪，本文利用回归损失函数的梯度来提取针对目标的显著性特征。

2.2 目标显著性感知特征

在预训练的分类网络中，每个卷积滤波器捕获特定的特征图案，并且所有的滤波器构建包含不同目标的特征空间。对于视觉跟踪任务，可以通过识别那些在目标区域活动的对象来获得具有与目标相关信息的过滤器。为此，本文将目标中心的图像的所有样本$X_{i,j}$回归到高斯图$Y(i,j)=\mathrm{e}^{-\frac{i^2+j^2}{2\sigma^2}}$，其中$(i,j)$是针对目标的偏移量，$\sigma$为样本标签的高斯函数标准差。为了计算效率，将该问题作为岭回归损失

$$L = \sum_{i,j} \|Y(i,j) - W * X_{i,j}\|^2 + \lambda \|W\|^2 \tag{5}$$

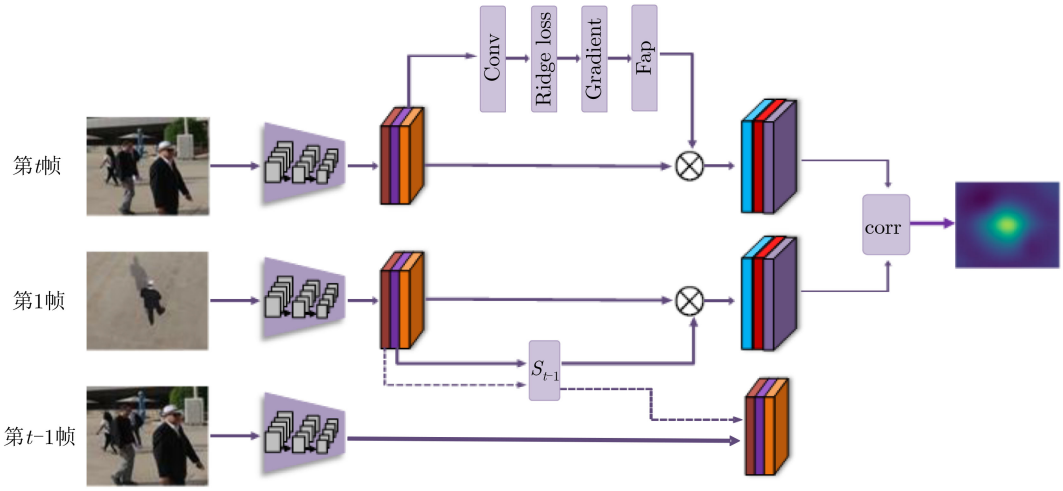


图 1 本文跟踪算法框架

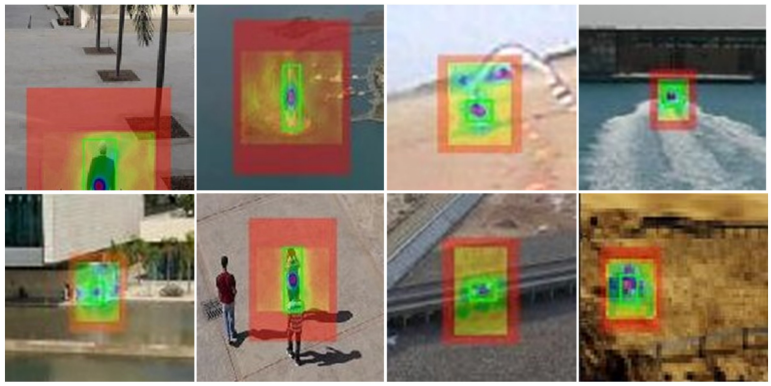


图 2 可视化跟踪效果图

其中，W为回归权重，每个滤波器的重要性可以根据其对拟合高斯图的贡献来计算，关于输入的特征X_{in}的损失函数L的推导可以根据链式法则和式(1)，回归损失的梯度由式(6)计算

$$\begin{aligned}\frac{\partial L}{\partial X_{\text{in}}}&=\sum_{i,j}\frac{\partial L}{\partial X_{\text{o}}(i,j)}\times\frac{\partial X_{\text{o}}(i,j)}{\partial X_{\text{in}}(i,j)}\\&=\sum_{i,j}2(Y(i,j)-X_{\text{o}}(i,j))\times W\end{aligned}\tag{6}$$

其中，X_{o}是输出预测，是将预测目标中心的图像的所有样本回归到高斯图$X_{\text{o}}(i,j)=\text{e}^{-\frac{i^2+j^2}{2\sigma^2}}$。从式(6)可知，$\dfrac{\partial L}{\partial X_{\text{in}}}$越小，预测的目标$X_{\text{o}}(i,j)$越接近真实的目标$Y(i,j)$，即输入的特征对正确跟踪目标的贡献越大，所以通过回归损失的梯度找到了能够区分目标和背景的目标感知滤波器。这不仅缓解了模型过度拟合的问题，而且减少了特征的数量。

2.3 目标干扰因子学习

对于单目标跟踪任务来说，丰富的目标姿势变化以及不同程度的背景干扰等跟踪难点都对提取具有辨别力的目标外观模型提出了更高的要求。然而，航拍视频中的目标是实时动态变化的，且基本上是俯视或侧俯视视角，仅使用第1帧作为固定模板限制了跟踪准确度的提高。本文在孪生网络的模板分支上加入动态学习干扰因子模块，引导在线学习第1帧的目标($\boldsymbol{O}_1$)与当前帧的前一帧目标($\boldsymbol{O}_{t-1}$)的相似性(如图3)，同时考虑目标的判别力特征和运动变化特征，在实时更新目标模板的同时避免模板漂移问题。

给定$\boldsymbol{X}$和$\boldsymbol{Y}$两个向量，我们的目的是寻找一个最优变换矩阵$\boldsymbol{R}$，使$\boldsymbol{X}$与$\boldsymbol{Y}$相似。使用线性回归方法有

$$\boldsymbol{R}=\arg\min_{\boldsymbol{T}}\|\boldsymbol{T}*\boldsymbol{X}-\boldsymbol{Y}\|^2+\lambda\|\boldsymbol{T}\|^2\tag{7}$$

在频域可快速求得$\boldsymbol{R}$

$$\boldsymbol{R}=\mathcal{F}^{-1}\left(\frac{\mathcal{F}^*(\boldsymbol{X})\odot\mathcal{F}(\boldsymbol{Y})}{\mathcal{F}^*(\boldsymbol{X})\odot\mathcal{F}(\boldsymbol{X})+\lambda}\right)\tag{8}$$

$\mathcal{F}$是离散傅里叶变换(DFT)，$\mathcal{F}^{-1}$是DFT的逆变换，上标*表示的是复共轭。

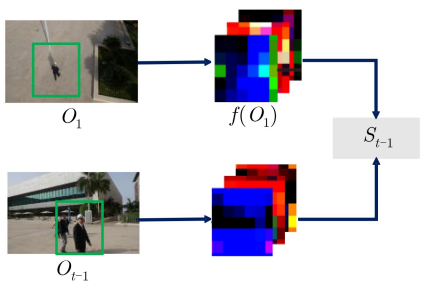


图 3 目标干扰因子学习

当跟踪到第$t-1$帧图像时，即可以得到目标$\boldsymbol{X}_{t-1}$。与简单地替换$\boldsymbol{Z}$和$\boldsymbol{X}_{t-1}$不同，本文是学习第1帧$\boldsymbol{Z}$到当前帧的前一帧$\boldsymbol{X}_{t-1}$的外观变化。本文假设外观变化在时间上是平滑的。因此，可以将这种变化应用于使$\boldsymbol{f}(\boldsymbol{Z})$相似于$\boldsymbol{f}(\boldsymbol{X}_{t-1})$，如式(7)所示。具体而言，本文使用线性回归计算得到目标干扰因子$\boldsymbol{S}_{t-1}$

$$\boldsymbol{S}_{t-1}=\arg\min_{s}\|\boldsymbol{S}*\boldsymbol{F}_1-\boldsymbol{F}_{t-1}\|^2+\lambda_s\|\boldsymbol{S}\|^2\tag{9}$$

其中，$\boldsymbol{F}_1=\boldsymbol{f}(\boldsymbol{Z})$, $\boldsymbol{F}_{t-1}=\boldsymbol{f}(\boldsymbol{X}_{t-1})$；$\lambda_{\boldsymbol{s}}$为正则化因子可从数据集学习得到。根据式(3)可知待搜索图像的目标显著性特征有

$$\boldsymbol{O}_t{}'=\boldsymbol{f}(\boldsymbol{O}_t;M_i)\tag{10}$$

从式(9)和式(10)可知目标跟踪响应结果可由式(11)计算

$$\text{res}=\text{corr}(\boldsymbol{S}_{t-1}*\boldsymbol{O}_1,\boldsymbol{O}_t{}')\tag{11}$$

2.4 算法实现

本文提出的跟踪算法流程如表1所示，通过反向传播分类神经网络得到的梯度来产生目标显著性特征，在频域上快速求解干扰因子。将目标显著性特征和在线学习干扰因子嵌入孪生网络中，从而实现有效可靠的自适应匹配跟踪。首先给定第1帧图像的目标位置和目标框，以第1帧目标为中心裁剪出比目标框大一点的区域作为目标模板，通过预训

表 1　本文跟踪算法流程

输入：

(1) 第1帧$\boldsymbol{Z}_1$：目标位置坐标$\boldsymbol{L}_1$和包围框$\boldsymbol{b}_1$

(2) 第$t-1$帧$\boldsymbol{X}_{t-1}$：目标位置坐标$\boldsymbol{L}_{t-1}$和包围框$\boldsymbol{b}_{t-1}$

(3) 第t帧$\boldsymbol{X}_t$(当前帧)

输出：

res：当前帧每个位置的相似度值

Function $(\boldsymbol{O}_t:\boldsymbol{O}_1:\boldsymbol{O}_{t-1})$= pretrain_feature $(\boldsymbol{X}_t:\boldsymbol{Z}_1:\boldsymbol{X}_{t-1})$

$(\boldsymbol{O}_t:\boldsymbol{O}_1:\boldsymbol{O}_{t-1})=\varphi(\boldsymbol{X}_t:\boldsymbol{Z}_1:\boldsymbol{X}_{t-1})$

end

Function $\boldsymbol{O}_t{}'$= Target_saliency_feature$(\boldsymbol{O}_t)$

$M_i=F_{\text{ap}}\left(\dfrac{\partial J}{\partial F_i}\right)$

$\boldsymbol{O}_t{}'=\boldsymbol{f}(\boldsymbol{O}_t;M_i)$

end

Function $\boldsymbol{S}_{t-1}$= get_disturbance_factor $(\boldsymbol{O}_1:\boldsymbol{O}_{t-1}:\lambda_s)$

$\boldsymbol{S}_{t-1}=\mathcal{F}^{-1}\left(\dfrac{\mathcal{F}^*(\boldsymbol{O}_1)\odot\mathcal{F}(\boldsymbol{O}_{t-1})}{\mathcal{F}^*(\boldsymbol{O}_1)\odot\mathcal{F}(\boldsymbol{O}_1)+\lambda_s}\right)$

end

Function res= detection $(\boldsymbol{S}_{t-1};\boldsymbol{O}_1;\boldsymbol{O}_t)$

$\text{res}=\text{corr}(\boldsymbol{S}_{t-1}*\boldsymbol{O}_1,\boldsymbol{O}_t{}')$

end

练网络来提取目标模板的特征。然后，开始进行第t帧跟踪，将第t帧图像以第$t-1$帧目标为中心进行裁剪，再将第$t-1$帧目标区域和第t帧待搜索区域分别经过预训练网络进行特征提取，此外还需将第t帧的预训练特征输入到目标显著性感知特征模块以提取到具有高度判别力的特征。然后将目标模板和第$t-1$帧提取的特征以及λ_s输入到目标干扰因子在线学习模块，来求得$\boldsymbol{S}_{t-1}$，再将目标干扰因子$\boldsymbol{S}_{t-1}$与目标模板特征卷积得到当前帧的模板。最后将第t帧的目标智能感知特征与第t帧模板进行相似度匹配得到目标的位置。

3 实验分析

3.1 数据集

实验部分采用航拍数据集UAV123[16]。UAV123建立了一个低空无人机视角下的高分辨率跟踪数据集,其中包括123个完整注释的空中视频序列，总帧数超过110 k帧。这些航拍视频具有更全面的目标跟踪干扰因素，其中包括低像素、相机运动、低分辨率、视角变化，光照变化、尺度变化、遮挡、形变、运动模糊、快速运动、平面内旋转、平面外旋转、出视野、背景干扰等航拍难点，由此可以看出用UAV123数据集来衡量一个目标跟踪算法的优劣更具有普适性和广泛性。

3.2 实验结果与分析

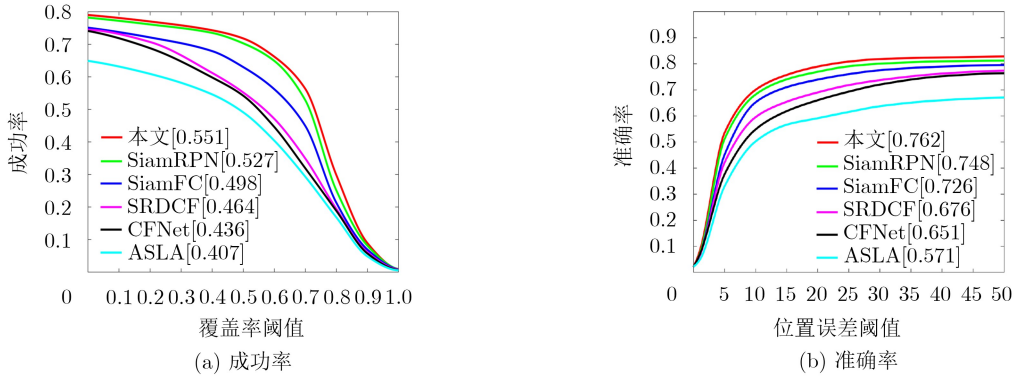
本文是在ILSVRC上离线训练的。采用随机梯度下降来最优化网络参数并设置权重衰减为0.0005，学习率以指数衰减方式从10^{-7}到10^{-9}，训练周期大约为50个周期且每次小批量训练样本数为8。首先初始化选择合适的λ_s，然后通过离线训练对其进行更新。实验平台为Matlab2016b，使用MatConvNet工具箱，使用的实验设备CPU为Intel Core i7-6700k 4 GHz, GPU为NVIDIA GeForce GTX1080，显存为8 GB。

3.2.1 定量分析

为验证本文算法的有效性，在广泛使用的无人机航拍数据集UAV123上进行实验。同时引入精确度(Precision)、成功率(Sucess)和速度(fps)这3个评价指标，将本文的算法跟踪结果与KCF[4], SiamFC[9], CSK[17], DSST[18], Struck[19], MEEM[20], MUSTER[21], SAMF[22], ASLA[23], CFNet[24], SRDCF[25]和SiamRPN[26]跟踪算法中选择跟踪效果最好的前5种算法来进行比较，进行定量分析。

图4是对比算法的成功率和精确度曲线图。由图4可以看出，本文算法相对于基准算法SiamFC[9]有很大的提升，成功率从49.8%提高到了55.1%，精确度也从72.6%提高到了76.2%。SiamRPN算法[26]包括用于特征提取的孪生网络和候选区域生成网络，其中候选区域生成网络包括分类和回归两条支路，而本文算法设计了一种新的特征通道回归损失函数来学习目标感知的深层特征，从而可以选择出选择最有效的卷积滤波器来生成待搜索目标显著性感知特征，还采用目标外观在线动态转换机制来实现有效可靠的自适应匹配跟踪，效果比SiamRPN算法[26]在成功率上提升了2.4%，在准确率上提升了1.4%。

表2展示了本文算法,KCF[4], Struck[19], MUSTER[21], SAMF[22], CFNet[24]和SRDCF[25]算法在10组代表性序列的成功率和准确率的比较(表2中数据为成功率/准确率)。从该表可以看出，相较于对比算法，本文算法在航拍场景下能够很好地跟踪目标。在航拍跟踪难点较多的building5, car15, truck2, uav4和wakeboard2序列中，本文算法的成功率均为第1，在penson21, car1_s和person3_s序列中，本文算法成功率也位列第2，在bike3, Building5, uav4, Wakeboard2和person3_s序列中，本文算法准确率位列第1，在car15, Person21, truck2, car1_s中本文算法准确率位列第2。表3统计了相关算法在10组视频序列的速度(fps)比较，虽然基于相关滤波的KCF[4]算法的速度确实很快，fps达到526.5帧/s，但是在视场大、目标小、背景复杂的航拍场景下，跟踪效果并不是很理想。本文算法速



(a) 成功率 (b) 准确率

图 4 成功率和准确率对比

表 2　部分视频的跟踪成功率和跟踪准确率比较(%)

序列	Attibutes	Struck	SAMF	MUSTER	KCF	SRDCF	CFNet	**本文**
bike3	LR POC	6.9/30.0	15.7/22.8	19.4/27.7	12.2/20.5	13.9/35.4	14.1/45.2	17.8/65.5
boat5	VC	16.6/10.6	74.7/61.8	85.1/67.7	23.2/9.5	48.6/89.7	36.0/17.2	38.7/37.6
building5	CM	99.3/99.3	96.9/98.6	97.3/98.9	89.0/99.7	97.5/98.9	21.6/61.6	99.8/99.8
car15	LR POC SOB	42.4/96	3.0/8.5	8.5/11.7	2.3/8.5	44.4/100.0	45.8/82.4	49.1/99.7
person21	LR POC VC SOB	31.2/43.9	0.6/9.4	51.3/79	0.6/5.7	30.8/82.9	18.6/47.2	28.7/73.9
truck2	LR POC	42.9/44.4	86.1/100	48.9/48.8	39.2/44.4	70.5/100.0	88.2/62.3	88.5/99.7
uav4	LR SOB	6.3/14.4	1.9/14.0	3.8/13.3	1.3/14.0	7.6/15.3	2.8/5.7	8.9/19.8
wakeboard2	VC CM	3.1/48.2	3.3/16.0	5.3/21.4	4.9/22.0	4.5/13.1	6.4/12.2	26.1/64.6
car1_s	POC OV VC CM	18.4/18.4	18.4/18.4	18.4/18.4	18.4/18.7	23.2/22.1	10.6/10.3	23.1/21.0
person3_s	POC OV CM	30.2/20.7	46.5/35	46.1/41.7	30.0/31.2	69.5/46.5	25.1/16.8	48.3/55.8

表 3　算法的速度(fps)比较

算法	Struck	SAMF	MUSTER	KCF	DCF	SRDCF	CFNet	本文
FPS	15.4	6.4	1.0	526.5	470.2	8.4	31.4	28.7

度虽然为28.7帧/s，但足以满足目标跟踪的实时性要求。

3.2.2　定性分析

为直接观察本文算法在航拍场景下的跟踪效果，选出具有航拍代表性属性的视频序列进行定性比较，主要分为5个类型：低分辨率、部分遮挡、出视野、视角变化、相似目标。由于航拍视频视场大而跟踪目标过小，为便于看清具体跟踪效果，全景视频帧侧边是以本文跟踪算法目标框为中心的裁剪图。通过测试视频效果图可以证明本文算法具有良好的准确性和鲁棒性。

(1) 低分辨率：由于无人机场景为俯瞰视角，所以在中高度距离拍摄的视频中跟踪目标往往为某个动态点目标，仅由几个像素组成，而且其大小与噪声非常相似，如图5所示。从图5(a)中可以看出，UAV7是无人机下拍摄的另外一台无人机的低分辨率视频序列，这是一种重要的空中跟踪场景，然而除本文算法外其他算法在后续帧都丢失了跟踪目标。在背景复杂场景下的bike3中，只有本文算法可以正确跟踪到目标。此外，在boat9的第96帧、第662帧和第1133帧跟踪目标出现了由小变大再变小的过程，相比之下本文算法可以更好更精确地框住目标。

(2) 部分遮挡：航拍视频下的遮挡不同于普通场景，是由于目标仅由几个像素构成，再出现遮挡使有效特征大大减少就会导致很难进行正确跟踪。在实际情况下，无人机拍摄的车辆在行驶过程中会受到树木以及标志牌的短时遮挡，如图6(a)、图6(c)，若没有在线更新模板(如SiamFC[9]等)肯定会造成目标丢失，导致再也跟踪不到正确目标。从car7, car9以及person4被遮挡后的第325帧，第1079帧以及第859帧可以看出本文算法可以鲁棒跟踪。

(3) 视角变化：空中无人机的任意方向运动会产生跟踪目标的视角变化，可能第i帧是目标的侧面，第j帧却是目标的背面。如图7所示，大部分算法在视角变化下还是可以正确跟踪目标的，但是可以看出本文算法能更有效更准确地框住目标。在bike1中出现了SiamFC[9]和Struck[19]的跟踪漂移，在person21中除本文算法外，由于出现相似目标其他算法都发生了一定的跟踪漂移。

(4)相似目标：和遮挡问题类似，航拍场景下的跟踪出现相似目标也是一大难点。在图8中展示的是具有代表性的测试帧的跟踪效果。在person21中还伴随视角变化、尺度变化和低分辨率，具有很大挑战性，在其他算法跟踪漂移的情况下，本文算法在第387帧目标被遮挡后能够快速调整正确跟踪。在car7中，在第251帧出现相似车辆时，MEEM[20]随即跟丢目标。person21跟踪情况类似，在后续帧中由于再出现遮挡，对比算法都开始陆续跟丢目标，而本文算法可以实现鲁棒效果。

4　结束语

针对航拍视频下的目标跟踪，本文提出了一种结合目标显著性和在线学习干扰变化的跟踪算法，实现了端到端的前馈式实时在线跟踪。与现有的跟踪算法相比，本文算法有3大优势。第一，本文提出目标感知特征学习，以缩小预训练的分类深度模型与特定航拍场景的目标跟踪之间的差距。其主要思想在于，由回归损失函数引起的梯度表明了相应

(a) UAV7

(b) truck3

(c) boat9

(d) bike3

本文算法 SiamRPN SiamFC MEEM MUSTER Struck KCF

图 5 视频序列测试图——低分辨率

(a) car7

(b) person4

(c) car9

本文算法 SiamRPN SiamFC MEEM MUSTER Struck KCF

图 6 视频序列测试图——部分遮挡

(a) bike1

(b) person21

(c) car16

(d) wakeboard3

本文算法　SiamRPN　SiamFC　MEEM　MUSTER　Struck　KCF

图 7 视频序列测试图——视角变化

(a) person21

(b) car7

(c) group1

本文算法　SiamRPN　SiamFC　MEEM　MUSTER　Struck　KCF

图 8 视频序列测试图——相似目标

滤波器在识别目标对象方面的重要性，因此通过从预先训练的CNN层中选择最有效的滤波器来学习航拍目标显著性深度特征。第二，本文模型具有可靠的在线适应能力，通过在线学习目标的干扰因子，来适应前景和背景的时间变化，提高了模型的鲁棒性，而不影响实时响应能力。第三，本文将目标显著性感知特征和在线学习干扰变化与孪生网络跟踪框架集成起来，在UAV123数据集上的广泛实验结果表明，该算法具有比较有竞争力的跟踪性能，同时还满足航拍视频的实时跟踪需求。

参 考 文 献

[1] TRILAKSONO B R, TRIADHITAMA R, ADIPRAWITA W, *et al*. Hardware-in-the-loop simulation for visual target tracking of octorotor UAV[J]. *Aircraft Engineering and Aerospace Technology*, 2011, 83(6): 407–419. doi: 10.1108/00022661111173289.

[2] 黄静琪, 胡琛, 孙山鹏, 等. 一种基于异步传感器网络的空间目标分布式跟踪方法[J]. 电子与信息学报, 2020, 42(5): 1132–1139. doi: 10.11999/JEIT190460.
HUANG Jingqi, HU Chen, SUN Shanpeng, *et al*. A distributed space target tracking algorithm based on asynchronous multi-sensor networks[J]. *Journal of Electronics & Information Technology*, 2020, 42(5): 1132–1139. doi: 10.11999/JEIT190460.

[3] KRISTAN M, MATAS J, LEONARDIS A, *et al*. A novel performance evaluation methodology for single-target trackers[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(11): 2137–2155. doi: 10.1109/TPAMI.2016.2516982.

[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. doi: 10.1109/TPAMI.2014.2345390.

[5] SUN Chong, WANG Dong, LU Huchuan, *et al*. Correlation tracking via joint discrimination and reliability learning[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 489–497. doi: 10.1109/CVPR.2018.00058.

[6] SUN Chong, WANG Dong, LU Huchuan, *et al*. Learning spatial-aware regressions for visual tracking[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8962–8970. doi: 10.1109/CVPR.2018.00934.

[7] QI Yuankai, ZHANG Shengping, QIN Lei, *et al*. Hedging deep features for visual tracking[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(5): 1116–1130. doi: 10.1109/TPAMI.2018.2828817.

[8] SONG Yibing, MA Chao, WU Xiaohe, *et al*. VITAL: Visual tracking via adversarial learning[C]. The 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, 2018: 8990–8999. doi: 10.1109/CVPR.2018.00937.

[9] BERTINETTO L, VALMADRE J, HENRIQUES J F, *et al*. Fully-convolutional Siamese networks for object tracking[C]. The 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016: 850–865. doi: 10.1007/978-3-319-48881-3_56.

[10] KRISTAN M, LEONARDIS A, MATAS J, *et al*. The sixth visual object tracking vot2018 challenge results[C]. Computer Vision ECCV 2018 Workshops, Munich, Germany, 2018: 3–53. doi: 10.1007/978-3-030-11009-3_1.

[11] CHEN Kai and TAO Wenbing. Once for all: A two-flow convolutional neural network for visual tracking[J]. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018, 28(12): 3377–3386. doi: 10.1109/TCSVT.2017.2757061.

[12] HELD D, THRUN S, SAVARESE S, *et al*. Learning to track at 100 fps with deep regression networks [C] The 14th European Conference on Computer Vision (ECCV) , Amsterdam，Netherlands, 2016, 9905: 749–765. doi: 10.1007/978-3-319-46448-0_45.

[13] 侯志强, 陈立琳, 余旺盛, 等. 基于双模板Siamese网络的鲁棒视觉跟踪算法[J]. 电子与信息学报, 2019, 41(9): 2247–2255. doi: 10.11999/JEIT181018.
HOU Zhiqiang, CHEN Lilin, YU Wangsheng, *et al*. Robust visual tracking algorithm based on Siamese network with dual templates[J]. *Journal of Electronics & Information Technology*, 2019, 41(9): 2247–2255. doi: 10.11999/JEIT181018.

[14] HUANG Chen, LUCEY S, RAMANAN D, *et al*. Learning policies for adaptive tracking with deep feature cascades[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 105–114. doi: 10.1109/ICCV.2017.21.

[15] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017: 618–626. doi: 10.1109/ICCV.2017.74.

[16] MUELLER M, SMITH N, and GHANEM B. A benchmark and simulator for UAV tracking[C]. The 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 2016: 445–461. doi: 10.1007/978-3-319-46448-0_27.

[17] HENRIQUES J F, CASEIRO R, MARTINS P, *et al*. Exploiting the circulant structure of tracking-by-detection with kernels[C]. The 12th European Conference on Computer Vision (ECCV), Florence, Italy, 2012: 702–715.

doi: 10.1007/978-3-642-33765-9_50.

[18] DANELLJAN M, HÄGER G, SHAHBAZ K, *et al*. Accurate scale estimation for robust visual tracking[C]. The British Machine Vision Conference (BMVC), Nottingham, UK, 2014: 65.1–65.11. doi: 10.5244/C.28.65.

[19] HARE S, GOLODETZ S, and SAFFARI A. Struck: Structured output tracking with kernels[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016, 38(10): 2096–2109. doi: 10.1109/TPAMI.2015.2509974.

[20] ZHANG Jianming, MA Shugao, and SCLAROFF S. MEEM: Robust tracking via multiple experts using entropy minimization[C]. The 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014: 188–203. doi: 10.1007/978-3-319-10599-4_13.

[21] HONG Zhibin, CHEN Zhe, WANG Chaohui, *et al*. Multi-store tracker (MUSTer): A cognitive psychology inspired approach to object tracking[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, USA, 2015: 749–758. doi: 10.1109/CVPR.2015.7298675.

[22] KRISTAN M, PFLUGFELDER R, LEONARDIS A, *et al*. The visual object tracking VOT2014 challenge results[C]. The 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 2014: 191–217. doi: 10.1007/978-3-319-16181-5_14.

[23] JIA Xu, LU Huchuan, and YANG M H. Visual tracking via adaptive structural local sparse appearance model[C]. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, USA, 2012: 1822–1829. doi: 10.1109/CVPR.2012.6247880.

[24] 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), Honolulu, USA, 2017: 5000–5008. doi: 10.1109/CVPR.2017.531.

[25] DANELLJAN M, HÄGER G, KHAN F S, *et al*. Learning spatially regularized correlation filters for visual tracking[C]. 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 4310–4318. doi: 10.1109/ICCV.2015.490.

[26] 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, USA, 2018: 8971–8980. doi: 10.1109/CVPR.2018.00935.

孙　锐：男，1976年生，教授，主要研究方向为计算机视觉与机器学习.

方林凤：女，1994年生，硕士生，研究方向为图像信息处理和计算机视觉.

梁启丽：女，1995年生，硕士生，研究方向为图像信息处理和计算机视觉.

张旭东：男，1966年生，教授，主要研究方向为智能信息处理.

责任编辑：马秀强