计算机科学
COMPUTER SCIENCE

# 基于隐空间匹配的无监督目标漂移校正及跟踪

范晓鹏, 彭力, 杨杰龙

引用本文

范晓鹏, 彭力, 杨杰龙. 基于隐空间匹配的无监督目标漂移校正及跟踪[J]. 计算机科学, 2024, 51(11): 166-173.

FAN Xiaopeng, PENG Li, YANG Jielong. Unsupervised Target Drift Correction and Tracking Based on Hidden Space Matching [J]. Computer Science, 2024, 51(11): 166-173.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900078

# 基于隐空间匹配的无监督目标漂移校正及跟踪

范晓鹏　彭　力　杨杰龙
江南大学物联网工程学院　江苏 无锡 214026
(fanxiaopeng2021@163.com)

**摘　要**　目标跟踪是计算机视觉领域的一个基础研究问题。随着跟踪技术的发展,现存的跟踪器主要存在两个挑战,即依赖于大量的数据标注信息和跟踪漂移,它们严重限制了跟踪器性能的提升。为了应对以上挑战,提出了无监督目标跟踪和隐空间匹配的方法。首先,通过可校正光流方法在前景中生成图像对;其次,利用生成的图像对从头开始训练孪生跟踪器;最后,使用隐空间匹配的方法,解决了跟踪器在目标形变较大、遮挡、出视野和漂移等情况下跟丢的问题。实验结果表明,算法 UHOT 的性能在多个数据集上有显著提升,在困难场景下展现出了较强的鲁棒性。与最新的无监督算法 SiamDF 相比,UHOT 在 VOT 数据集上取得了 8% 的增益,与最新的监督孪生跟踪器相当。

**关键词**:无监督;滑动窗口;隐空间;模板匹配;目标跟踪
**中图分类号**　TP391.4

## Unsupervised Target Drift Correction and Tracking Based on Hidden Space Matching

FAN Xiaopeng,PENG Li and YANG Jielong
School of Internet of Things Engineering,Jiangnan University,Wuxi,Jiangsu 214026,China

**Abstract**　Object tracking is a basic research issue in the field of computer vision. With the development oftracking technology, existing trackers mainly have two challenges,namely relying on a large amount of data annotation information and tracking drift, which seriously limits the improvement of tracker performance. In order to overcome the above challenges,unsupervised target tracking and hidden space matching methods are proposed. Firstly,image pairs are generated in the foreground via a correctable optical flow method. Secondly,the generated image pairs are utilized to train the siamese tracker from scratch. Finally,the hidden space matching method is used to solve the problem of losing track when the target deforms greatly,is occluded,goes out of the field of view and drifting. Experimental results show that the algorithm UHOT significantly improves on multiple datasets and demonstrates strong robustness in difficult scenarios. Compared with the latest unsupervised algorithm SiamDF,UHOT gaines 8% gain on the VOT dataset,comparable to state-of-the-art supervised siamese trackers.

**Keywords**　Unsupervised,Sliding window,Hidden space,Template matching,Object tracking

## 1　引言

目标跟踪是计算机视觉中一项基本任务,为众多应用程序提供了重要支持。具体来说,目标跟踪是智能监控、无人驾驶、智能交通、虚拟现实等应用程序的关键环节,旨在给定视频初始帧中的目标,寻找后续帧中目标所在的位置。但跟踪问题仍然面临着依赖人工标注数据集,在低像素、遮挡、出视野情况下的跟丢等问题。为了让目标跟踪不再依赖标注信息,部分研究仅使用视频序列初始帧的目标信息进行跟踪。Wang 等通过前后向跟踪监督多帧一致性损失训练了一个基于判别相关过滤器的跟踪 UDT[1];$S^2$ SiamFC[2] 通过随机裁剪单帧的方式生成伪框,通过对抗性掩蔽构建模板-搜索对。但是上述方法的跟踪性能很大程度上依赖于在线更新,如果没有在线更新,这些无监督的训练跟踪器就无法处理具有显著变化的物体。一些研究人员提出将光流[3]应用到神经网络上,计算目标周围像素的光流向量,迭代更新实现对目标的跟踪。光流方法虽然不依赖在线更新,但是不适合光照强度变化大的目标跟踪,在目标形变大、遮挡和出视野的情况下面临着严峻的挑战,从而影响模型的全局优化。因此,本文提出了可校正光流来寻找移动的目标,结合时序信息分析空间移动信息,对视频序列进行时序优化。为了解决在目标形变大、遮挡和出视野的情况下目标跟丢无法找回的问题,提出了隐空间匹配来找回目标,采用运动估计的方法,将模板和搜索区域进行关键特征映射,通过计算隐空间的相似度,找出与模板对应的目标区域。本文工作的主要贡献总结如下:

1)通过可校正光流发现移动目标,运用滑动窗口筛选和

到稿日期:2023-09-14　返修日期:2024-01-28
基金项目:国家自然科学基金青年科学基金(62106082)
This work was supported by the Young Scientists Fund of the National Natural Science Foundation of China(62106082).
通信作者:彭力(pengli@jiangnan.edu.cn)

补帧，生成可靠的候选框实现无监督学习。

2）采用运动估计方法将图像向隐空间映射，利用隐空间对应关系解决目标丢失的问题。

3）在多个数据集上实现了精度和鲁棒性的提升。

## 2 相关工作

### 2.1 无监督目标跟踪

由于视频序列注释成本较高，因此无监督跟踪成为解决问题的主流方案。UDT 基于判别式相关滤波器 DCFNet 训练了一个前后向轨迹周期一致性损失的跟踪器。$S^2$ SiamFC 提出了一种基于孪生网络的无监督训练框架，通过对比学习在单帧中构建模板-搜索对进行训练，在精度上逐渐逼近有监督的跟踪器[4-8]。Zheng 等提出了一种无监督的跟踪方法 USOT[9]，在第一阶段从单帧开始进行初始训练，然后在更长的时间跨度上进行周期训练。PUL[10]通过对比学习识别前景背景，进一步考虑噪声损失，获得了优异的性能。上述工作都极大地依赖前后帧的一致性损失，然而中间帧跟踪结果的好坏严重影响跟踪性能。与这些工作不同，本方法考虑每一帧的跟踪结果，获得了优越的跟踪性能。

### 2.2 模板匹配方法

基于相关滤波器的算法[11]是一种简单实用的目标跟踪算法，主要思想是先对每个目标进行检测，生成一系列候选框，然后采用卡尔曼滤波对目标的位置和速度进行估计，并根据卡尔曼滤波的结果来匹配目标。虽然该方法速度快，但在复杂场景下可能出现漏检。Meanshift[12]算法是一种基于特征直方图的目标跟踪算法，主要思想是在当前帧中，通过目标模板的特征直方图计算最佳匹配位置，但在周围存在相似干扰情况下的效果很差。最近的基于孪生网络的目标跟踪方法的主要思想是通过两个共享网络分别提取目标模板和搜索区域，然后将两个特征向量送入一个相似度评估模块，计算它们之间的相似度，获取最佳匹配位置。但由于缺乏在线更新，该方法无法适应遮挡、目标变化较大的场景。SiamMask[13]在分类和回归的基础上，增加了 Mask 分支来实现目标分割[14]，可以适应变化大的场景，有很好的鲁棒性和精度，但是跟踪的速度面临挑战。本文提出的隐空间目标匹配很好地解决了上述问题，将模板和搜索区域通过运动估计的方法向隐空间映射，计算隐空间的相似度，不仅能达到很好的精度和鲁棒性，还能达到实时的效果。

## 3 本文方法

在本文方法中，无监督训练跟踪器由两部分组成，网络整体方案如图 1 所示。在 3.1 节中，本文方法使用可校正光流对视频序列进行处理寻找目标，并生成候选框序列，继而生成图像对，将其输入孪生跟踪器进行无监督主干训练。此方法可以解决跟踪器需要依赖注释的问题。在 3.2 节中，通过隐空间匹配的方式寻找目标，继续训练孪生跟踪器，这是在更长的时间跨度内执行的，能够解决目标丢失、遮挡等复杂跟踪问题。

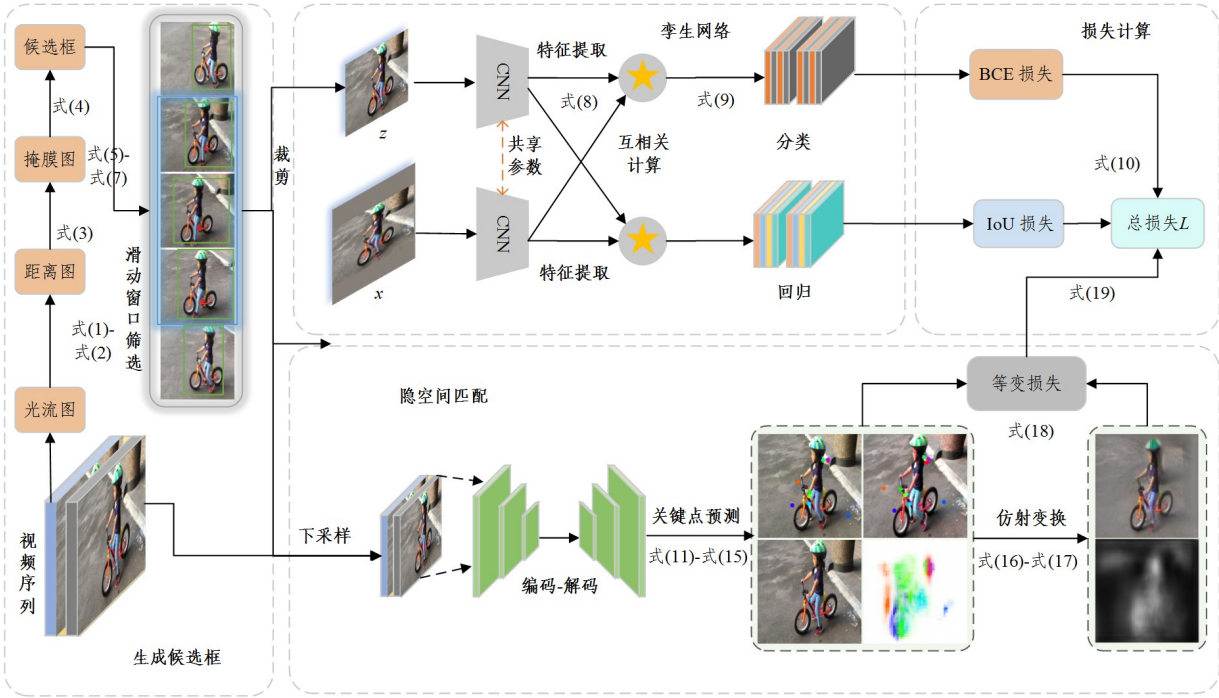


图 1 无监督跟踪整体框架图特征

Fig. 1 Overall framework of unsupervised training tracker

### 3.1 无监督生成候选框进行训练

不同于 S2SiamFC 随机裁剪对象生成图像对标记视频序列，本文方法使用可校正光流对视频序列进行处理，可以获得相对准确的运动目标。由于物体的运动轨迹趋于平滑状态，因此可以通过滑动窗口插值补帧的方法获得可靠的跟踪序列。

光流是一种基于像素运动的方法，通常以两张图像作为输入，预测每个像素的位移。利用这种特性，可以捕捉连续帧之间的物体运动，找出运动的前景目标。由于光流生成候选框受相机运动、遮挡等影响，因此采用滑动窗口的方法筛选可靠的候选框，并对缺失的帧进行插值补帧处理。

#### 3.1.1 候选框生成

为了应对跟踪器依赖标签的挑战，采用光流对前景运动目标进行提取，如图 2 所示。对于任意一个包含 $L$ 个相同大小 $W\times H$ 连续帧的视频($W$ 为宽，$H$ 为高)，使用现成的无监督 ARFlow 算法从 $t$ 帧和 $t+i$ 帧中计算光流图 $F_t$，$i$ 为间隔的帧数。假设光流图中某一像素点 $p$ 的光流向量为$(u,v)$，$u$ 表示水平方向的分量，$v$ 表示垂直方向的分量，然后将像素点 $p$ 的 $F_t^p$ 转换为距离 $D_t^p$：

$$F_t^p=\sqrt{(u_t^p)^2+(v_t^p)^2} \tag{1}$$

$$D_t^p=\sqrt{(\Delta u_t^p)^2+(\Delta v_t^p)^2} \tag{2}$$

其中，$\Delta u_t^p$ 和 $\Delta v_t^p$ 分别表示 $u$ 和 $v$ 在 $p$ 点的差值。将 $D_t^p$ 进行二值化，则像素点 $p$ 的掩模图 $M_t^p$ 如下：

$$M_t^p=\begin{cases}1, & \text{若 } D_t^p\geqslant\alpha\cdot max(D_t)+(1-\alpha)\cdot mean(D_t)\\ 0, & \text{否则}\end{cases} \tag{3}$$

其中，$D_t$ 表示距离图，而 $\alpha\in(0,1)$是一个超参数，空间维度内的最大值和平均值分别用 $max$ 和 $mean$ 表示。

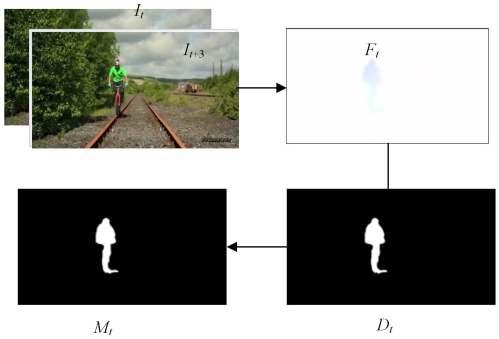


图 2 基于光流生成的目标候选框

Fig. 2 Target candidate box generated based on optical flow

掩模图 $M_t^p=1$ 的像素区域表示原图 $I_t$ 对应物体运动的区域。为了进一步从这些候选区域中过滤出不可靠的区域，本文方法将这些区域的矩形边界作为初始候选框，并根据大小和位置对这些候选框进行评分。由于中心偏差，较大的候选框在图像的中间应该有更高的质量分数。设候选框 $B_t=(x_0,y_0,x_1,y_1)$表示一个候选框的左上角和右下角的坐标。框 $B_t$ 的评估分数 $E(B_t)$被定义为：

$$E(B_t)=(x_1-x_0)(y_1-y_0)+\beta\cdot min(x_0,W-x_1)min(y_0,H-y_1) \tag{4}$$

其中，$W$ 和 $H$ 分别表示图片的宽和高，$\beta$ 是一个超参数，得分高的候选框被选择作为 $I_t$ 帧的最终候选框 $B_t$。将视频中所有被选择的候选框 $B_t$ 的集合表示为 $B=\{B_t\mid 1\leqslant t\leqslant L\}$。

#### 3.1.2 滑动窗口筛选和插值补帧

生成的候选框集合 $B$ 可能包含因摄像机抖动、遮挡等而产生的噪声框，为了去除不可靠的候选框，采用滑动窗口的方法对候选框集合 $B$ 进行过滤并生成新候选框。从候选框集合中选择 $n$ 个连续帧，为了衡量运动目标的平滑程度，考虑中心坐标偏移 $S_{dev}$和多帧的重叠面积 $S_{IoU}$两个指标。假设每个候选框的中心坐标是$(x^1,y^1),(x^2,y^2),\cdots,(x^n,y^n)$，那么滑动窗口内第 $i$ 帧的中心坐标的偏移量的得分 $S_{dev}^i$如下：

$$S_{dev}^i=\frac{\sum_n((x^i-x^{mean})^2+(y^i-y^{mean})^2)}{n} \tag{5}$$

其中，$x^{mean}$ 和 $y^{mean}$ 分别表示滑动窗口内的 $x$ 和 $y$ 的平均值。第 $i$ 帧的重叠面积得分为 $S_{IoU}^i$：

$$S_{IoU}^i=\frac{(S^i\cap S^1)\cup(S^i\cap S^2)\cup(S^i\cap S^n)}{S^1\cup S^2\cup S^i\cup\cdots\cup S^n} \tag{6}$$

其中，$n$ 表示滑动窗口的帧数，一般取 $n=5$。$S^i$ 表示第 $i$ 帧的面积，那么第 $i$ 个候选框的平滑得分 $S_{sw}^i$为：

$$S_{sw}^i=S_{IoU}^i-\rho S_{dev}^i \tag{7}$$

其中，$\rho$ 是一个超参数，为了鼓励一个平滑的轨迹，在 $S_{dev}^i$上为距离惩罚设置 $\rho>1$。对于没有被平滑机制选择为候选框的帧，使用线性插值，根据上述筛选后的相邻候选框生成伪框，如图 3 所示。将生成的候选框通过裁剪等操作生成一个新的帧，称为滑窗帧 $z$，该帧对应的未处理的帧称为原始帧 $x$。

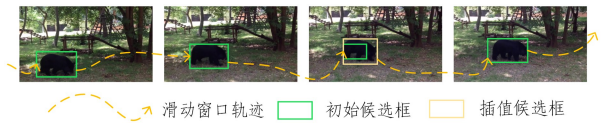


图 3 通过线性插值的方式给相邻的候选框生成伪框

Fig. 3 Generating pseudo boxes for adjacent candidate boxes by linear interpolation

#### 3.1.3 无监督孪生网络跟踪

将上述经过滑动窗口筛选和插值补帧得到的图像对送入孪生网络进行主干训练。SiamFC 是一个用于对象跟踪任务的全卷积孪生网络框架，其核心思想是解决跟踪作为一个互相关和相似度学习问题[15]。将上述生成的 $z$ 和 $x$ 输入孪生网络中进行训练，可以在原始帧 $x$ 内得到相应目标区域。相似度函数计算式为：

$$f(z,x)=\varphi(z)*\varphi(x)+\mu \tag{8}$$

其中，$\varphi(z)$和 $\varphi(x)$分别表示通过 ResNet[16] 网络对 $z$ 和 $x$ 进行特征提取，$*$ 是互相关运算，$\mu$ 表示一个偏差项。将式(11)得到的相似度 $f$ 使用 SoftMax 转换为概率分布：

$$c=\frac{\exp(f)}{\sum\exp(f)} \tag{9}$$

如图 1 孪生网络模块所示，输出响应图有两个分支：一个用于前景和背景分类；另一个输出回归响应图，用于表示从中心位置到边界框 4 个边的距离。上述初始帧 $x$ 根据目标区域进行二分类，用于计算分类损失，标签为 $g$。其中 $g=1$ 为候选框目标所在的区域，那么损失函数 $L_{naive}$ 就是回归损失和分类损失的和：

$$L_{naive}=L_{reg}(f,g)+\lambda_1 L_{cls}(c,g) \tag{10}$$

其中，$L_{reg}$ 和 $L_{cls}$ 分别为 IoU 损失[17] 和二元交叉熵 BCE 损失[18]，$\lambda_1$ 是一个权重参数。

### 3.2 隐空间匹配方案

为了应对图像序列的形变、遮挡、漂移和出视野等变化较大的情况，提出了隐空间匹配。隐空间匹配主要由两个模块组成：关键点运动估计模块和局部仿射变换模块。通过关键点运动估计可以找到模板帧和搜索帧的共性特征，将其作为关键点；局部仿射变换对关键点周围特征向隐空间映射，在隐空间进行模板匹配。该方案可以解决遮挡、形变大的情况下的跟踪漂移问题，做到准确的跟踪。

隐空间匹配如图 1 所示。运动估计模块的目的是预测搜索帧 $S\in\mathcal{R}^{3\times H\times W}$到模板帧 $T\in\mathcal{R}^{3\times H\times W}$的密集运动场。搜索

帧 $S$ 和模板帧 $T$ 来自同一个视频序列,且它们之间的间隔不相差 $i$ 帧。密集运动场是一个 $S$ 到 $T$ 的映射函数 $\gamma_{T\leftarrow S}:\mathcal{R}^2\rightarrow\mathcal{R}^2$,这里采用了反向光流 $\gamma$,可以使用双线性采样[19]以可微的方式有效地实现反向映射。由于存在遮挡、形变大和出视野等情况,因此运动估计模块不能直接预测 $\gamma_{S\leftarrow T}$ 和 $\gamma_{T\leftarrow S}$。假设存在一个隐空间坐标系 $X$,在模板帧和搜索帧之间生成了隐空间的帧 $R$,该帧通过视频序列生成,同时具备两者特征的共性,可以独立估计两个转换:从隐空间帧到模板帧 $\gamma_{T\leftarrow R}$ 和从隐空间到搜索帧 $\gamma_{S\leftarrow R}$,这样就可以独立处理模板帧和搜索帧。

3.2.1 关键点运动估计模块

关键点运动估计模块通过无监督方法对视频序列进行学习,可以完成对运动对象的捕捉,标记模板帧和搜索帧的共性特征。受 Monkey-Net[20]启发,训练期间每次将模板帧与固定间隔帧内的搜索帧输入,构建一个潜在表示的帧来训练模型,该帧即为隐空间的帧。每次输入的模板帧和搜索帧会组成一个模板对,通过目标对象提取和编码解码,计算模板对的相关性,可以得到前景运动目标的关键点位移。

运动估计模块估计搜索帧到模板帧的反向光流 $\gamma_{T\leftarrow S}$。假设存在一个隐空间 $R$,估计 $\gamma_{T\leftarrow S}$ 包含 $\gamma_{R\leftarrow S}$ 和 $\gamma_{T\leftarrow R}$。此外,给定一个坐标系 $X$,估计关键点附近的每个变换 $\gamma_{X\leftarrow R}$,考虑在 $k$ 个关键点 $p_1\cdots p_k$ 的一阶泰勒展开式,$p_1\cdots p_k$ 表示 $R$ 中的关键点坐标。为了简单起见,关键点在隐空间中都用 $p$ 来表示,而关键点位置在 $X$ 坐标系中用 $z$ 表示。可以得到:

$$\gamma_{X\leftarrow R}(p)=\gamma_{X\leftarrow R}(p_k)+\left(\frac{d}{dp}\gamma_{X\leftarrow R}(p)\,|\,p=p_k\right)(p-p_k)+o(\|p-p_k\|) \tag{11}$$

其中,函数 $\gamma_{X\leftarrow R}$ 每个关键点 $p_k$ 中的值和位置可以用雅可比矩阵[21]来表示:

$$\gamma_{X\leftarrow R}(p)\simeq\left\{\left\{\gamma_{X\leftarrow R}(p_1),\frac{d}{dp}\gamma_{X\leftarrow R}(p)\,\Big|_{p=p_1}\right\},\cdots,\left\{\gamma_{X\leftarrow R}(p_k),\frac{d}{dp}\gamma_{X\leftarrow R}(p)\,|_{p=p_k}\right\}\right\} \tag{12}$$

假设 $\gamma_{X\leftarrow R}$ 在每个关键点附近都是一对一映射的,那么 $\gamma_{R\leftarrow X}=\gamma_{X\leftarrow R}^{-1}$。为了估计搜索帧关键点附近的 $\gamma_{T\leftarrow S}$,首先估计搜索帧 $S$ 中点 $z_k$ 附近的变换 $\gamma_{R\leftarrow S}$,然后估计 $R$ 中 $p_k$ 附近的变换 $\gamma_{T\leftarrow R}$,最后得到 $\gamma_{T\leftarrow S}$:

$$\gamma_{T\leftarrow S}=\gamma_{T\leftarrow R}\cdot\gamma_{R\leftarrow S}=\gamma_{T\leftarrow R}\cdot\gamma_{S\leftarrow R}^{-1} \tag{13}$$

将式(13)进行一阶泰勒展开,得到:

$$\gamma_{T\leftarrow S}(z)=\gamma_{T\leftarrow R}(p_k)+J_k(z-\gamma_{S\leftarrow R}(p_k)) \tag{14}$$

$$J_k=\left(\frac{d}{dp}\gamma_{T\leftarrow R}(p)\,\Big|_{p=p_k}\right)\left(\frac{d}{dp}\gamma_{S\leftarrow R}(p)\,\Big|_{p=p_k}\right)^{-1} \tag{15}$$

实际上,式(14)中的 $\gamma_{T\leftarrow R}(p_k)$ 和 $\gamma_{S\leftarrow R}(p_k)$ 均由关键点预测器得出。使用标准的 U-Net[22]架构来估计 $k$ 个热图,每个关键点位置的预测都是使用平均操作来估计的,每个关键点和热力图一一对应,解码器的最后一层使用 Softmax 函数来预测关键点置信度。

对于模板帧和搜索帧,关键点预测网络还输出 4 个额外的通道。从这些通道中,将对应的关键点置信度图作为权重来计算空间加权平均值,计算出式中 $\frac{d}{dp}\gamma_{T\leftarrow R}(p)\,|_{p=p_k}$ 和 $\frac{d}{dp}\gamma_{S\leftarrow R}(p)\,|_{p=p_k}$ 的值。

3.2.2 局部仿射变换

使用卷积网络,通过模板帧和关键点中的 $\gamma_{T\leftarrow S}(z)$ 的 Taylor 近似集来估计 $\hat{\gamma}_{T\leftarrow S}$。需要注意的是,虽然 $\hat{\gamma}_{T\leftarrow S}$ 在搜索帧的每个像素与模板帧的对应位置一一映射,但是在边缘和纹理部分,映射的效果不佳,这个问题让网络很难从模板帧去估计 $\hat{\gamma}_{T\leftarrow S}$。为了让 $\hat{\gamma}_{T\leftarrow S}$ 有大致对齐的效果,我们根据式(13)对模板帧 $T$ 进行局部变换,得到 $k$ 个变换后的图像 $T^1$,$T^2$,…,$T^k$。

对于每个关键点 $p_k$,将计算变换过程的密集运动网络作为热图 $H_k$,$H_k(z)$ 表示以 $\gamma_{T\leftarrow R}(p_k)$ 和 $\gamma_{S\leftarrow R}(p_k)$ 为中心的两个热图的差值。热图 $H_k(z)$ 的计算方法如下:

$$H_k(z)=\exp\left(\frac{(\gamma_{S\leftarrow R}(p_k)-z)^2}{\sigma}\right)-\exp\left(\frac{(\gamma_{T\leftarrow R}(p_k)-z)^2}{\sigma}\right) \tag{16}$$

其中,$\sigma$ 为超参数,取 $\sigma=0.01$。热图 $H_k$ 和转换后的图像 $T^1$,$T^2$,…,$T^k$ 被送入 U-Net 处理,并使用一个受 Monkey-Net 启发的 part-based 网络对其进行估计,通过对每个部分进行独立的变换,将不同部分的特征进行连接、融合或加权求和,实现目标的综合表示,从而预测目标的真实位置。这样,网络就能够综合使用各个部分的信息,并全面地捕捉目标的整体形态和姿态。通过这种 part-based 网络的设计,跟踪器能够更好地应对目标的形变、遮挡以及出视野等问题。假设一个物体由 $K$ 个刚性部分组成,每个部分都根据式(14)移动,估计 $K+1$ 个映射 $M_k(k=0,\cdots,K)$ 表示哪个地方发生了局部变换。最终的密集运动预测为:

$$\hat{\gamma}_{T\leftarrow S}(z)=M_0 z+\sum_{k=1}^{K}M_k(\gamma_{T\leftarrow R}(p_k)+J_k(z-\gamma_{S\leftarrow R}(p_k))) \tag{17}$$

其中,$M_0 z$ 是隐空间中非移动的部分,如背景。

3.2.3 损失函数

关键点预测器整个过程在无监督条件下运行,在训练期间不需要任何注释,这可能会导致性能不稳定。等方差约束[23]是无监督关键点检测的最重要因素之一,它强制模型经过几何变换存在相关的关键点,使用薄板样条变换,被用于无监督自然图像变形的关键点检测。由于运动估计不仅可以预测关键点,还可以预测关键点周围的局部变换,因此,将等方差损失 $L_{eq}$ 扩展到对局部仿射变换约束。

$$L_{\text{eq}}=|\gamma_{S\leftarrow R}^{k}-\tilde{\gamma}\gamma_{S\leftarrow R}^{k}| \tag{18}$$

其中,$\tilde{S}$ 是由 $S$ 经过仿射变换得到的,$\tilde{\gamma}$ 是由一些随机变换组成的。因此,最后的损失和如下:

$$L=L_{\text{naive}}+\lambda_2 L_{\text{eq}} \tag{19}$$

其中,$\lambda_2$ 是一个超参数,一般取 $\lambda_2<1$。

## 4 实验

本章介绍了无监督跟踪器在多个基准数据集上的结果,并与最先进的跟踪算法进行了比较。通过广泛的消融实验来分析跟踪器的有效性。

### 4.1 数据准备

本文算法是基于 Pytorch0.7.1 深度学习框架实现的,操作系统为 Ubuntu16.04,16 GB 内存,CPU 为英特尔 i7-8700,

显卡是 NVIDIAGeForceGTX1080，8 GB 显存。实验选用 GOT-10K 数据集进行离线训练，该数据集大约包含 10000 个视频序列，未使用标注信息，视频中的每一帧都有几个视觉属性，包括遮挡、照明变化、运动变化、大小变化、相机运动或出视野。整个训练过程经历了 30 个阶段，其中隐空间匹配只在最后 25 个 epochs 内进行。前 5 个 epochs 的学习率从 $2.5\times10^{-3}$ 开始到 $5\times10^{-3}$，而剩下的 epochs 中采用的学习率从 $5\times10^{-3}$ 到 $2\times10^{-5}$ 呈指数下降。

训练标签是由数据集 GOT-10k 上使用的无监督光流模型生成的，数据采样策略类似于 USOT，训练集的真实标签在训练中是不可用的。从一个视频中采样多个模板帧和搜索帧，在具有较大时间间隔的区域内进行周期训练。模板帧被裁剪为 $127\times127$，将其作为前后向跟踪的参考，搜索帧大小为 $255\times25$。

### 4.2 实验结果分析

与最先进的跟踪方法的比较是在多个数据集上进行的，包括 VOT2016，VOT2018[24] 和 OTB100[25]。结果表明，UHOT 的表现超过了最先进的无监督跟踪器。

OTB2015：当前单目标跟踪领域应用最广泛的数据集之一，它包含了 100 组视频序列，具备光照变化(IV)、尺度变化(SV)、遮挡(OCC)、形变(DEF)、运动模糊(MB)、快动作(FM)、平面内旋转(IPR)、平面外旋转(OPR)、离开视野(OV)、背景复杂(BC)以及分辨率低(LR)共 11 项困难属性，采用重叠成功率和欧氏距离精度来评估跟踪器的性能。UHOT 在各项指标达上到了不错的效果，如图 4 所示。此外，UHOT 的效果比有监督踪器 SiamFC 更优。

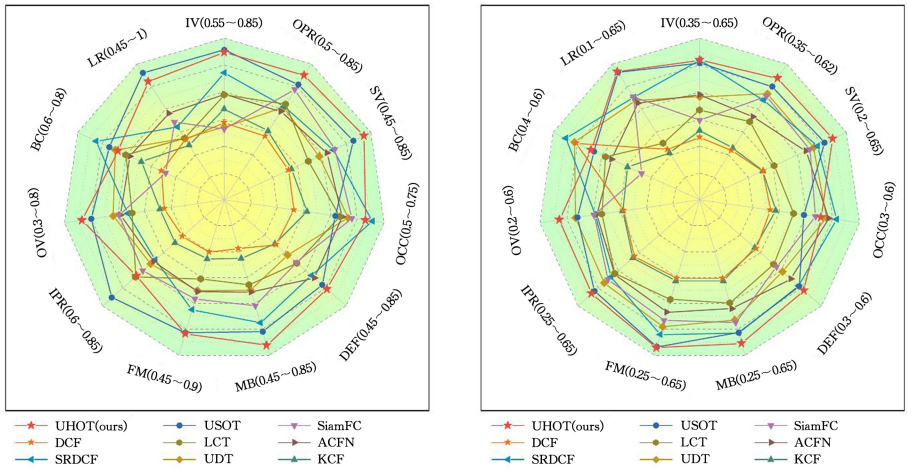


图 4 各方法在 OTB2015 数据集上的精度和成功率雷达图

Fig. 4 Radar chart of precision and success rate of different methods on OTB2015 dataset

图 5 展示了几种算法在 OTB2015 数据集上 6 个复杂序列的跟踪结果可视化。将 UDT+，SiamFC，USOT 与 UHOT 进行对比，除了 SiamFC，其他都是无监督跟踪器。实验结果表明，本文算法在这些序列上能够准确地定位目标。多个序列的跟踪难点在于尺度变化、快速运动、遮挡以及光照变化。整个跟踪过程中，目标都处于背景复杂的条件下且周围有相似物体干扰；而 UHOT 能够在每一帧都成功定位目标，得益于隐空间匹配，增强了对目标重要特征的表达，实现了从结构到语义多方面的物体表征。本文算法预测框能够准确完整地锁定目标，进一步体现了 UHOT 在应对复杂环境时具备较强的鲁棒性。

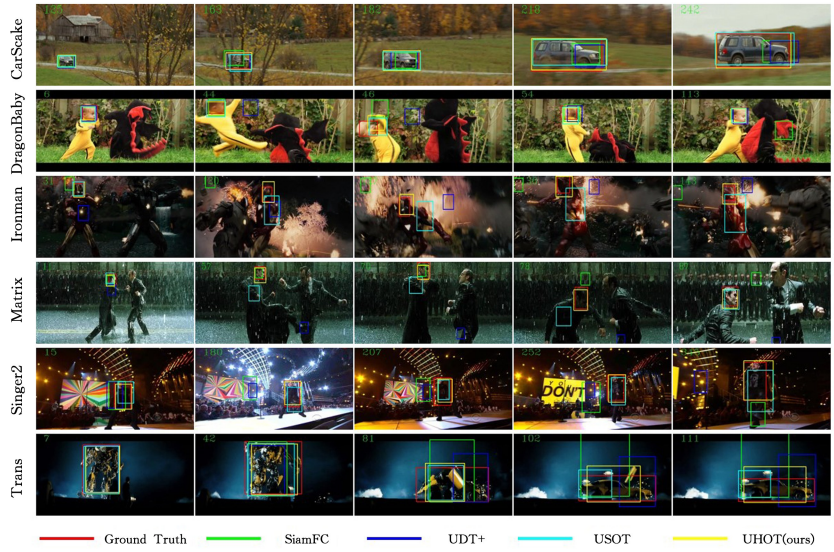


图 5 OTB2015 数据集上一些序列的跟踪结果

Fig. 5 Tracking results of some sequences on OTB2015 dataset

VOT2016：数据集包含 60 个视频序列，该标准包括精度（$A$）、鲁棒性（$R$）和期望平均重叠（$EAO$）。由于 VOT 的目标是短期的视觉跟踪，如果跟踪失败，将涉及一个重新初始化机制。表 1 中的结果表明，UHOT 显著优于最先进的无监督跟踪器，比性能最好的无监督跟踪器 USOT 的 $EAO$ 提升了 5%，且性能优于基于部分监督算法 SiamFC 和 AutoTrack。最先进的无监督方法 SiamDF，虽然在鲁棒性上实现了领先的性能，但依赖于大规模数据集，而 UHOT 只利用同一图像的训练对，并以自我监督和离线的方式进行训练。

表 1 12 种跟踪器在 VOT2016 上的评估

Table 1 Evaluation of 12 trackers on VOT2016

| 跟踪器 | 年份 | 无监督 | $A\uparrow$ | $R\downarrow$ | $EAO\uparrow$ |
|---|---|---|---|---|---|
| UDT | 2019 | √ | 0.539 | 0.475 | 0.225 |
| UDT+ | 2019 | √ | 0.534 | 0.655 | 0.200 |
| $S^2$SiamFC | 2020 | √ | 0.493 | 0.639 | 0.215 |
| USOT | 2021 | √ | 0.593 | 0.336 | 0.351 |
| ResPUL | 2021 | √ | 0.554 | 0.405 | 0.263 |
| AlexPUL | 2021 | √ | 0.548 | 0.545 | 0.219 |
| LUDT+ | 2022 | √ | 0.57 | 0.331 | 0.299 |
| SiamDF | 2023 | √ | 0.52 | 0.238 | 0.339 |
| SiamFC | 2016 | × | 0.53 | 0.461 | 0.235 |
| SiamDW | 2019 | × | 0.54 | 0.38 | 0.303 |
| AutoTrack | 2020 | × | 0.517 | 0.26 | 0.271 |
| UHOT | Our | √ | 0.603 | 0.363 | 0.371 |

VOT2018：数据集同样由 60 个视频序列组成，但 VOT2018 用更困难的序列替换了 VOT2016 数据集里的 10 个简单的片段。VOT2018 的评估标准也与 VOT 系列相同，但测试的序列更具挑战性。比较结果如表 2 所列。UHOT 取得了比 SiamFC，SiamDW[26] 和 AutoTrack[27] 更具竞争力的结果。虽然 ATOM[28] 等方法实现了最先进的性能，但其主要受益于标注信息和多个大规模数据集。与 USOT 相比，UHOT 的 $EAO$ 提高了 4%。

表 2 13 种跟踪器在 VOT2018 上的评估

Table 2 Evaluation of 13 trackers on VOT2018

| 跟踪器 | 年份 | 无监督 | $A\uparrow$ | $R\downarrow$ | $EAO\uparrow$ |
|---|---|---|---|---|---|
| UDT | 2019 | √ | 0.472 | 0.932 | 0.129 |
| UDT+ | 2019 | √ | 0.650 | 0.670 | 0.276 |
| LUDT+ | 2020 | √ | 0.490 | 0.412 | 0.230 |
| $S^2$SiamFC | 2020 | √ | 0.463 | 0.782 | 0.180 |
| USOT | 2021 | √ | 0.564 | 0.435 | 0.290 |
| AlexPUL | 2021 | √ | 0.515 | 0.693 | 0.182 |
| ResPUL | 2021 | √ | 0.516 | 0.660 | 0.203 |
| SiamDF | 2023 | √ | 0.505 | 0.450 | 0.250 |
| SiamFC | 2016 | × | 0.503 | 0.585 | 0.188 |
| SiamDW | 2019 | × | 0.500 | 0.490 | 0.234 |
| ATOM | 2019 | × | 0.590 | 0.204 | 0.401 |
| AutoTrack | 2020 | × | 0.484 | 0.391 | 0.200 |
| UHOT | Our | √ | 0.584 | 0.438 | 0.303 |

### 4.3 消融实验

1）光流生成候选框：UHOT 在 OTB2015 基准测试集上进行实验，研究候选框生成策略。与随机裁剪生成候选框相比，提出的光流候选框序列进行训练后精度有显著的提升。使用两种不同的策略在 OTB2015 数据集上进行测试，结果如表 3 所列。

2）滑动窗口插值补帧：为了评估滑动窗口筛选机制和插值补帧的贡献，对训练的不同阶段进行了广泛的消融研究。值得注意的是，该模块在光流生成候选框的操作后进行，以弥补部分场景下光流处理能力弱的短板。这个设计的动机是充分利用视频序列的时空连贯性，减少对光流生成候选框的依赖。为了验证这一点，采用 3 种策略在 OTB2015 基准测试集上进行实验，3 种策略分别为直接使用光流生成的视频序列、滑动窗口筛选后的视频序列和插值补帧后的视频序列。实验结果如表 3 所列，结果表明，经过滑动窗口机制筛选和插值补帧后，跟踪器精度显著提升。

表 3 候选框选取策略在 OTB2015 数据集上的结果

Table 3 Results of candidate box selection strategy on OTB2015 dataset

| 候选框筛选和补帧 | $Auc$ | $Pre$ |
|---|---|---|
| 光流生成 | 0.582 | 0.785 |
| 光流生成+滑动窗口 | 0.591 | 0.791 |
| 光流生成+滑动窗口+插值补帧 | 0.603 | 0.822 |

3）模板帧和搜索帧的选择：为了充分说明模板帧和搜索帧的输入对跟踪结果的影响，对训练的输入进行了消融实验。采用两种方式进行输入，分别是从图像序列中随机抽取和从固定间隔帧内随机抽取，结果如表 4 所列。实验结果表明，从固定间隔帧内随机抽取的效果明显好于从整个图像序列中随机抽取，因为目标变化和背景变化非常迅速，从图像序列中随机抽取，搜索帧和模板帧目标和背景相关性很小，隐空间匹配效果差；而从固定间隔帧内随机抽取一定程度上限制了移动目标变化的时间和空间，模板帧和搜索帧相关性较大，隐空间匹配效果好，精度得以提升。

表 4 模板帧和搜索帧选取策略在 OTB2015 数据集上的结果

Table 4 Results of template frame and search frame selection strategies on OTB2015 dataset

| 模板帧和搜索帧选取策略 | $Auc$ | $Pre$ |
|---|---|---|
| 图像序列中随机抽取 | 0.563 | 0.761 |
| 固定间隔帧内随机抽取 | 0.601 | 0.819 |

4）隐空间匹配：为了更好地理解隐空间匹配，证明该方法可以很好地处理视频序列遮挡、出视野等情况，将所提出的候选框生成、初始孪生网络和隐空间匹配进行比较，分别表示为 GB，SF 和 TM。用孪生网络训练和隐空间匹配进行对比实验，结果如表 5 所列，仅使用隐空间匹配的方法失去了孪生网络的主干训练，精度有所下降，将两者结合训练，可以获得一个很好的结果。观察表 6，UHOT 在 OTB2015 数据集上对遮挡（OCC）、离开视野（OV）、快动作（FM）3 个困难指标进行评估。在隐空间匹配加持下，精度（$A$）、鲁棒性（$R$）和预期平均重叠（$EAO$）都得到有效提高，证明 UHOT 可以很好地解决遮挡、出视野、形变大等复杂情况的跟踪问题。

表 5 训练策略在 VOT2016 数据集上的结果

Table 5 Results of training strategies on VOT2016 dataset

| 方法 | $A\uparrow$ | $R\downarrow$ | $EAO\uparrow$ |
|---|---|---|---|
| GB+SF | 0.597 | 0.389 | 0.354 |
| GB+TM | 0.594 | 0.351 | 0.363 |
| GB+SF+TM | 0.603 | 0.363 | 0.371 |

表 6 隐空间匹配在 OTB2015 复杂场景的结果

Table 6 Results of hidden space matching in OTB2015 complex scenarios

| 跟踪器 | *Auc/Pre* | | |
|---|---|---|---|
| | OCC | OV | FM |
| GB+SF | 0.508/0.700 | 0.465/0.630 | 0.561/0.738 |
| GB+SF+TM | 0.528/0.721 | 0.552/0.838 | 0.631/0.745 |

图 6 展示了跟踪器在多个序列上出现了跟丢后的结果。对于 Human8 序列，跟踪器在 30 帧出现了跟丢的情况，在第 32 帧后续帧再次准确定位目标，这得益于隐空间匹配。Jump 序列是一个跟踪难度较高序列，跟踪难点在于形变、模糊和快速移动。所有的跟踪器在 37 帧几乎都跟丢了目标，在后续的两帧中，UHOT 找回了目标并持续跟踪，而其他跟踪器在跟丢之后输出几乎没有变化。Soccer 序列存在遮挡和模糊等特点，所有的跟踪器在遮挡变化下都丢失了目标，但后续只有 UHOT 找回了目标，解决了目标丢失的问题。

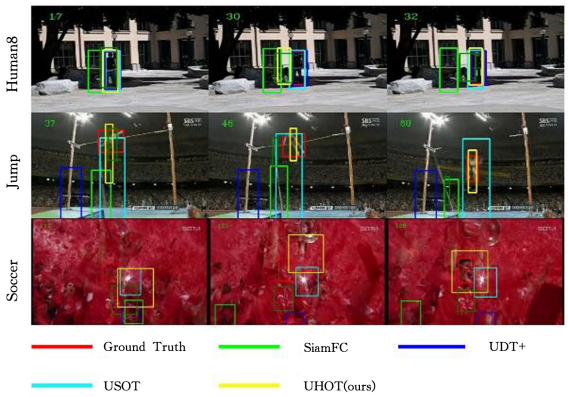


图 6 跟踪器在多个序列上跟丢后的表现

Fig. 6 Performance of the tracker after tracking loss on multiple sequences

**结束语** 本文提出了一种新的无监督跟踪框架 UHOT。首先，利用可校正光流生成的图像对训练一个初始的孪生网络跟踪器，该方法只用图像对进行离线训练，而不需要任何注释。然后，继续结合关键点特征和局部仿射变换的方法，生成目标在隐空间的表示，进行目标融合匹配，在更长的时间跨度内训练跟踪器，解决了跟踪器在遮挡、形变大、出视野和漂移情况下丢失目标的问题。大量的实验表明，所提出的无监督跟踪器与最新的无监督跟踪器性能相当。最后，考虑到模型的精度，在模型的大小上做了取舍，因此模型极度依赖于设备，未来可以把该无监督模型向轻量化的方向发展。

## 参 考 文 献

[1] WANG N, SONG Y B, MA C, et al. Unsupervised deep tracking [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1308-1317.

[2] CHONG H S, MA Y J, CHEN J C, et al. S2siamfc: Self-supervised fully convolutional siamese network for visual tracking [C]// Proceedings of the 28th ACM International Conference on Multimedia. 2020: 1948-1957.

[3] LIU L, ZHANG J N, HE R F, et al. Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 6489-6498.

[4] LUCA B, JACK V, JOAO F, et al. Fully-convolutional siamese networks for object tracking [C] // Computer Vision — ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8—10 and 15—16, 2016, Proceedings, Part II 14. Springer, 2016: 850-865.

[5] CAI H Y, LAN L, ZHANG J, et al. Siamdf: Tracking training data-free siamese tracker[M]. Neural Networks, 2023.

[6] LI H J, PENG L. High-speed tracking algorithm based on negative sample mining and feature fusion [J]. Control and Decision Making, 2023, 38(9): 2554-2562.

[7] SUN K W, WANG Z H, LIU H, et al. Maximum Overlap Single Target Tracking Algorithm Based on Attention Mechanism. [J]. Computer Science, 2023, 50(S1): 397-401.

[8] ZENG Z H, LUO H L. Cross-dataset Learning Combining Multi-object Tracking and Human Pose Estimation[J]. Computer Science, 2023, 50(S1): 512-518.

[9] ZHENG J L, MA C, PENG H W, et al. Learning to track objects from unlabeled videos[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13546-13555.

[10] WU Q Q, WAN J, ANTONI B C. Progressive unsupervised learning for visual object tracking [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 2993-3002.

[11] ZHANG J, LIU Y, LIU H, et al. Learning local-global multiple correlation filters for robust visual tracking with Kalman filter redetection[J]. Sensors, 2021, 21(4): 1129.

[12] LE N, RATHOUR V S, YAMAZAKI K, et al. Deep reinforcement learning in computer vision: a comprehensive survey[J]. Artificial Intelligence Review, 2022, 55(4): 2733-2819.

[13] WANG Q, LI Z, LUCA B, et al. Fast online object tracking and segmentation: A unifying approach [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1328-1338.

[14] CHENG Y M, LI L L, XU Y Y, et al. Segment and track anything[J]. arXiv: 2305.06558, 2023.

[15] LI B, YAN J J, WU W, et al. High performance visual tracking with siamese region proposal network[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8971-8980.

[16] TARG S S, DIOGO A, KEVIN L. Resnet in resnet: Generalizing residual architectures[J]. arXiv: 1603.08029, 2016.

[17] YU J H, JIANG Y N, WANG Z Y, et al. Unitbox: An advanced object detection network[C]// Proceedings of the 24th ACM International Conference on Multimedia. 2016: 516-520.

[18] HO Y S, SAMUEL W. The real-worldweight cross-entropy loss function: Modeling the costs of mislabeling[J]. IEEE Access, 2019, 8: 4806-4813.

[19] JADERBERG M, SIMON K, ZISSERMAN A, et al. Spatial transformer networks. [J]. arXiv: 1506.02025, 2015.

[20] JOOST V A, ANITHA K, MARC A R, et al. Transformation-based models of video sequences[J]. arXiv: 1701.08435, 2017.

[21] TOMAS J,ANKUSH G,HAKAN B,et al. Unsupervised learning of object landmarks through conditional image generation [C]// Advances in Neural Information Processing Systems 31. 2018.

[22] OLAF R,PHILIPP F,THOMAS B. U-net:Convolutional networks for biomedical image segmentation[C]// Medical Image Computing and Computer-Assisted Intervention-MICCAI 2015: 18th Interna= tional Conference, Munich, Germany, October 5 -9,2015,Proceedings,Part III 18. Springer,2015:234-241.

[23] ALIAK S,STEPHANE L,SERGEY T,et al. First order motion model for image animation[C]// Advances in Neural Information Processing Systems 32. 2019.

[24] MATEJ K,ALES L,JIRI M,et al. The sixth visual object tracking vot2018 challenge results[C]// Proceedings of the European Conference on Computer Vision(ECCV). 2018.

[25] WU Y,LI J W,YANG M H. Online object tracking:A benchmark[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013:2411-2418.

[26] ZHANG Z P,PENG H W. Deeper and wider siamese networks for real-time visual tracking[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4591-4600.

[27] LI Y M,FU C H,DING F Q,et al. Autotrack:Towards high-performance visual tracking for uav with automatic spatio-temporal regularization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020.

[28] MARTIN D,GOUTAM B,FAHAD S K,et al. Atom:Accurate tracking by overlap maximization [C] // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:4660-4669.

**FAN Xiaopeng**, born in 1998, postgraduate. His main research interests include object tracking and pose estimation.

**PENG Li**, born in 1967, Ph.D, professor. His main research interests include computer vision and pattern recognition.

(责任编辑:何杨)