(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112489088 A
(43)申请公布日 2021.03.12

(21)申请号 202011473954.6

(22)申请日 2020.12.15

(71)申请人 东北大学
地址 110819 辽宁省沈阳市和平区文化路3号巷11号

(72)发明人 于瑞云 杨骞 王开开 李张杰

(74)专利代理机构 沈阳东大知识产权代理有限公司 21109
代理人 吴琼

(51)Int.Cl.
G06T 7/246(2017.01)
G06T 7/238(2017.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

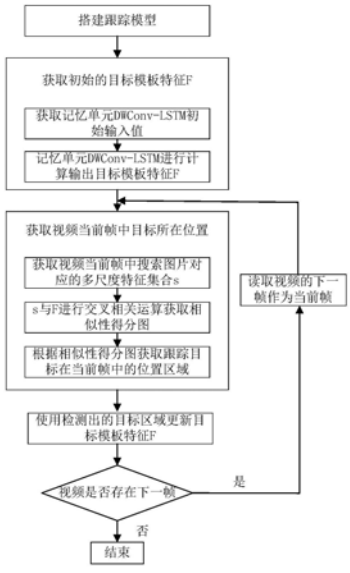
权利要求书1页 说明书5页 附图5页

(54)发明名称

一种基于记忆单元的孪生网络视觉跟踪方法

(57)摘要

一种基于记忆单元的孪生网络视觉跟踪方法，属于目标跟踪技术领域，包括：步骤1、搭建跟踪模型；步骤2、获取初始的目标模板特征；步骤3、获取跟踪目标在视频当前帧中的对应位置；步骤4、根据步骤3中找出的当前帧中跟踪目标所在位置裁剪出目标所在区域作为目标模板，将该目标模板输入跟踪模型的目标模板分支获取新的目标模板特征；读取视频的下一帧作为当前帧，转入步骤3进行下一轮迭代，直至读取完视频中的所有帧结束迭代。本发明方法能够有效应对视觉跟踪过程中的遮挡、背景的混杂和目标的形态剧烈变化等问题，提升了跟踪模型面对复杂环境时的跟踪鲁棒性。



CN 112489088 A

1.一种基于记忆单元的孪生网络视觉跟踪方法,其特征在于,具体步骤如下:

步骤1:搭建跟踪模型;

跟踪模型分为两个分支:分别是目标模板分支和搜索图像分支,目标模板分支由骨干网络和DWConv-LSTM记忆单元两个模块组成,搜索图像分支由骨干网络组成,两个分支的骨干网络是共享权值的基于残差模块搭建的孪生网络;

步骤2:获取初始的目标模板特征F;

步骤3:获取跟踪目标在视频当前帧中的对应位置;

步骤4:根据步骤3中找出的当前帧中跟踪目标所在位置裁剪出目标所在区域作为目标模板,将该目标模板输入跟踪模型的目标模板分支获取新的目标模板特征F;读取视频的下一帧作为当前帧,转入步骤3进行下一轮迭代,直至读取完视频中的所有帧结束迭代。

2.根据权利要求1所述的基于记忆单元的孪生网络视觉跟踪方法,其特征在于,所述步骤1中,所述DWConv-LSTM记忆单元为使用深度可分离卷积代替长短期记忆网络LSTM中的全连接结构建立的,使得记忆单元输出的目标模板特征既包含时序上的变化信息又包含空间上的特征信息。

3.根据权利要求1所述的基于记忆单元的孪生网络视觉跟踪方法,其特征在于,所述步骤2获取初始的目标模板特征F,具体操作步骤为:

步骤2.1:初始的目标模板图像经过目标模板分支的骨干网络提取出特征e_0,e_0通过3×3的卷积和1×1的卷积运算得到DWConv-LSTM记忆单元中初始的细胞单元c_0,e_0通过由另一组3×3的卷积和1×1的卷积组成的支路进行卷积运算得到DWConv-LSTM记忆单元中初始隐藏层状态h_0;

步骤2.2:将c_0,h_0和e_0输入到DWConv-LSTM记忆单元中,获取初始的目标模板特征F。

4.根据权利要求1所述的基于记忆单元的孪生网络视觉跟踪方法,其特征在于,所述步骤3获取跟踪目标在视频当前帧中的对应位置,具体操作步骤为:

步骤3.1:从视频当前帧中裁剪出搜索图像,获取多尺度的搜索图像对应的特征;

从视频中的当前帧中裁剪出搜索图像S_t,建立与S_t对应的多尺度搜索图像集合$S=\{S_t^{-1},S_t^0,S_t^1\}$,集合S中的多尺度搜索图像作为一个批次经过跟踪模型的搜索图像分支得到多尺度搜索图像对应的特征集合$s=\{s_t^{-1},s_t^0,s_t^1\}$;

步骤3.2:获取相似性得分图;

当前时刻的目标模板特征F与步骤3.1中得到的多尺度搜索图像对应的特征$\{s_t^{-1},s_t^0,s_t^1\}$根据如下公式进行交叉相关运算,得到相似性得分图集合$r=\{r_t^{-1},r_t^0,r_t^1\}$,式中(*)代表卷积算子;

$$r_t^i = F * s_t^i, i = -1,0,1$$

步骤3.3:根据相似性得分图获取跟踪目标对应位置;

对集合r中的每一个相似性得分图进行上采样,得到上采样后的相似性得分图集合$R=\{R_t^{-1},R_t^0,R_t^1\}$,在集合R中找出最大值所在的上采样后的相似性得分图,将其标记为R_t,对R_t中所有值进行比较,获取值最大的K个响应值点,这K个响应值点求平均值得到响应值点d,在视频当前帧中找到d点对应位置即为所寻找的目标所在的位置。

一种基于记忆单元的孪生网络视觉跟踪方法

技术领域

[0001] 本发明属于目标跟踪技术领域，具体涉及一种基于记忆单元的孪生网络视觉跟踪方法。

背景技术

[0002] 在计算机视觉技术中目标跟踪技术扮演着举足轻重的角色，广泛的应用在智慧交通、安防、体育、医疗、机器人导航和人机交互等领域，具有巨大的商业价值。视觉跟踪要完成的任务是在一个图像序列中选择一个感兴趣的区域作为跟踪目标，在接下来的连续若干图像帧中获得准确的目标位置、具体形态和目标的运动轨迹等信息。从技术发展角度来说，视觉跟踪技术的研究可以分为三个阶段，在第一阶段，以卡尔曼滤波、均值滤波、粒子滤波及光流法为代表的经典跟踪方法；在第二阶段，以TLD模型为代表的基于检测的视觉跟踪方法，以CSK算法为代表的相关滤波类视觉跟踪方法；在第三个阶段，基于深度学习的视觉跟踪方法。但是在视觉跟踪任务中，只能使用第一张图像的目标标注信息，因而在训练的过程中缺乏足够的先验知识来保证跟踪模型的精度。此外视觉跟踪问题还面临着光照的变化、跟踪目标的严重遮挡、背景的混杂、目标的形态剧烈变化以及运动模糊等挑战。

[0003] 基于孪生网络的视觉跟踪模型将跟踪问题转化成了图片块相似度匹配问题。孪生网络跟踪模型将目标模板图像和视频中当前帧对应的搜索图像作为网络的输入向量，通常搜索图像的面积较大些。两个张量经过骨干网络得到各自的特征，将目标模板对应的特征当作卷积核与搜索图片对应的特征进行交叉相关运算，最终得到一个相似性得分图，相似性得分图中最大值点所在位置就是视频当前帧中目标所在的位置。

[0004] 然而在传统的孪生网络跟踪模型中，仅仅使用初始帧作为目标模板，模型在离线训练完成后就不再对模型的网络参数和目标模板进行更新。网络参数不更新意味着模型在遇到未见过的场景或目标时的跟踪性能会出现巨大滑坡，目标模板不更新将会在视频序列中的目标发生剧烈的外观变化或受到严重遮挡等情况时产生跟踪漂移问题，这些都会导致模型鲁棒性和跟踪精度的下降。

发明内容

[0005] 针对传统的孪生网络跟踪模型中存在的问题，本发明从目标模板的更新角度入手，利用DWConv-LSTM记忆单元从时间和空间两个角度去解决目标模板的更新问题，能够在提高模型的鲁棒性的同时获得较好的跟踪准确率，具有相当的使用价值。

[0006] 本发明的技术方案是，在跟踪模型中骨干网络采用基于残差的孪生网络，模型有两个分支，上边支路为目标模板分支，下边支路为搜索图像分支。在目标模板分支中，目标模板通过骨干网络和记忆单元两个模块获得目标的鲁棒特征；在搜索图像分支中，为适应跟踪过程中目标尺度变化，对搜索图像做多尺度处理得到三种不同尺度的搜索图像，这些搜索图像经过骨干网络提取特征。三种不同尺度的搜索图像对应的特征分别与目标模板对应的特征进行交叉相关运算，最终得到三个相似性得分图，找到三者中最优的特征图，目标

的预测位置由该特征图上值最大的K个响应值的平均值决定。根据检出的目标位置获取新的目标模板，将获取的目标模板输入跟踪模型的目标模板分支，利用DWConv-LSTM记忆单元学习目标模板在时间和空间上的变化。

[0007] 本发明的一种基于记忆单元的孪生网络视觉跟踪方法，具体步骤如下：

[0008] 步骤1：搭建跟踪模型；

[0009] 跟踪模型分为两个分支：分别是目标模板分支和搜索图像分支，目标模板分支由骨干网络和DWConv-LSTM记忆单元两个模块组成，搜索图像分支由骨干网络组成，两个分支的骨干网络是共享权值的基于残差模块搭建的孪生网络；

[0010] DWConv-LSTM记忆单元实质上为融合了深度可分离卷积操作的长短期记忆网络LSTM，传统的LSTM可以很好的描述时序信息，但内部采用全连接结构，这样使得LSTM学习到大量对描述时序无用的信息，并且不能像卷积操作那样刻画目标的空间特征信息；本发明采用深度可分离卷积代替LSTM中的全连接结构，使得记忆单元输出的目标模板特征既包含时序上的变化信息又包含空间上的特征信息；

[0011] 步骤2：获取初始的目标模板特征F；

[0012] 步骤2.1：初始的目标模板图像经过目标模板分支的骨干网络提取出特征e_0，e_0通过3×3的卷积和1×1的卷积运算得到DWConv-LSTM记忆单元中初始的细胞单元c_0，e_0通过由另一组3×3的卷积和1×1的卷积组成的支路进行卷积运算得到DWConv-LSTM记忆单元中初始隐藏层状态h_0；

[0013] 步骤2.2：将c_0，h_0和e_0输入到DWConv-LSTM记忆单元中，获取初始的目标模板特征F；

[0014] 步骤3：获取跟踪目标在视频当前帧中的对应位置；

[0015] 步骤3.1：从视频当前帧中裁剪出搜索图像，获取多尺度的搜索图像对应的特征；

[0016] 假设当前帧是视频的第t帧(t＝1,2,3,…,n)，从视频中的当前帧中裁剪出搜索图像S_t，建立与S_t对应的多尺度搜索图像集合$S=\{S_t^{-1},S_t^0,S_t^1\}$，集合S中的多尺度搜索图像作为一个批次经过跟踪模型的搜索图像分支得到多尺度搜索图像对应的特征集合$s=\{s_t^{-1},s_t^0,s_t^1\}$；

[0017] 步骤3.2：获取相似性得分图；

[0018] 当前时刻的目标模板特征F与步骤3.1中得到的多尺度搜索图像对应的特征$\{s_t^{-1},s_t^0,s_t^1\}$根据公式(1)进行交叉相关运算，得到相似性得分图集合$r=\{r_t^{-1},r_t^0,r_t^1\}$，式中(*)代表卷积算子；

[0019] $$r_t^i = F * s_t^i, i=-1,0,1 \quad (1)$$

[0020] 步骤3.3：根据相似性得分图获取跟踪目标对应位置；

[0021] 对集合r中的每一个相似性得分图进行上采样，得到上采样后的相似性得分图集合$R=\{R_t^{-1},R_t^0,R_t^1\}$，在集合R中找出最大值所在的上采样后的相似性得分图，将其标记为R_t，对R_t中所有值进行比较，获取值最大的K个响应值点，这K个响应值点求平均值得到响应值点d，在视频当前帧中找到d点对应位置即为所寻找的目标所在的位置；

[0022] 步骤4：根据步骤3中找出的当前帧中跟踪目标所在位置裁剪出目标所在区域作为

目标模板，将该目标模板输入跟踪模型的目标模板分支获取新的目标模板特征F；读取视频的下一帧作为当前帧，转入步骤3进行下一轮迭代，直至读取完视频中的所有帧结束迭代。

[0023] 上述一种基于记忆单元的孪生网络视觉跟踪方法，其中：

[0024] 所述步骤3中，初始时视频的第1帧作为当前帧，即第一轮迭代时t＝1，之后在由步骤4转入步骤3进行下一轮迭代时，视频的下一帧作为当前帧，即此时t＝t+1。

[0025] 所述步骤4中，在更新目标模板特征F的过程中，DWConv-LSTM记忆单元会以上一次更新过程中DWConv-LSTM记忆单元输出的细胞单元c_{t-1}和隐层状态h_{t-1}，以及目标模板分支骨干网络提取出来的特征e_t作为输入，来更新细胞单元和隐层状态，并获取目标模板特征F。

[0026] 本发明的收益为：

[0027] 本发明在跟踪模型的目标模板分支添加的基于DWConv-LSTM的记忆单元可以学习目标在时间序列上的外观变化趋势，同时利用卷积网络也保证了目标在空间上的稳定性，能够有效应对视觉跟踪过程中的遮挡、背景的混杂和目标的形态剧烈变化等问题，提升了跟踪模型面对复杂环境时的跟踪鲁棒性，同时在记忆单元中使用深度可分离卷积进一步的加速了记忆单元，使得跟踪模型能够保证实时性，有重要应用价值。

附图说明

[0028] 图1本发明的基于记忆单元的孪生网络视觉跟踪方法流程图。

[0029] 图2本发明的跟踪模型整体架构图

[0030] 图3本发明的DWConv-LSTM记忆单元的基本网络结构图。

[0031] 图4本发明的目标模板图像的裁剪示例图。

[0032] 图5本发明获取记忆单元初始的细胞单元和隐层状态过程示意图。

[0033] 图6本发明在视频帧中裁剪多尺度的搜索图像示例图。

[0034] 图7、8本发明具体实施例效果图。

具体实施方式

[0035] 下面结合附图对本发明的具体实施方式做详细说明。

[0036] 本实施方式的方法，算法实现采用TensorFlow深度学习框架，操作系统为Ubuntu16.04LTS。

[0037] 如图1所示，一种基于记忆单元的孪生网络视觉跟踪方法，具体步骤如下：

[0038] 步骤1：搭建跟踪模型；

[0039] 如图2所示，跟踪模型分为目标模板分支和搜索图像分支两个分支，目标模板分支由骨干网络和DWConv-LSTM记忆单元两个模块组成，搜索图像分支由骨干网络组成，两个分支的骨干网络是共享权值的基于残差模块搭建的孪生网络；

[0040] DWConv-LSTM记忆单元的网络结构如图3所示，DWConv-LSTM记忆单元内部主要包含遗忘门f_t、输入门i_t和输出门o_t三个门控单元和一个细胞单元c_t，各个门的计算分别按公式(2)(3)(4)进行，其中W_i、W_f和W_o为权重矩阵，b_i、b_f和b_o为偏移量，e_t为目标模板分支骨干网络提取出来的特征，h_{t-1}为上一时刻隐层状态，(*)代表卷积算子，σ代表sigmoid激活函数；

[0041] $$i_t = \sigma\left(W_i * \begin{bmatrix} e_t \\ h_{t-1} \end{bmatrix} + b_i\right) \quad (2)$$

[0042] $$f_t = \sigma\left(W_f * \begin{bmatrix} e_t \\ h_{t-1} \end{bmatrix} + b_f\right) \quad (3)$$

[0043] $$o_t = \sigma\left(W_o * \begin{bmatrix} e_t \\ h_{t-1} \end{bmatrix} + b_o\right) \quad (4)$$

[0044] 细胞单元c_t的更新，隐层状态h_t的更新及目标模板特征F的获取分别依据公式(5)(6)(7)进行计算；其中，tanh代表tanh激活函数，c_{t-1}为上一时刻细胞单元，(*)代表卷积算子，W_c和W_t为权重矩阵，b_c为偏移量，h_{t-1}为上一时刻隐层状态；

[0045] $$c_t = f_t * c_{t-1} + i_t * \tanh\left(W_c * \begin{bmatrix} e_t \\ h_{t-1} \end{bmatrix} + b_c\right) \quad (5)$$

[0046] $h_t = o_t * \tanh(c_t)$ (6)

[0047] $F = W_t * h_t$ (7)

[0048] 各个门的卷积运算都是使用深度可分离卷积实现，有助于捕获特征之间的空间关系并减少网络参数量从而加快前向推理的速度。

[0049] 步骤2：获取初始的目标模板特征向量F；

[0050] 步骤2.1：根据给定的待检测跟踪目标的宽w和高h，按照公式(8)计算出跟踪目标所在区域大小Z，其中，p表示扩充长度，按照公式(9)计算得到。如图4所示，在原图中以给定的待检测跟踪目标中心位置(cx,cy)为中心裁剪出边长为$\sqrt{Z}$的正方形区域即为模板图像。再将该模板图像高-宽尺寸调整成127×127大小，得到初始的目标模板。如图5所示，初始的目标模板经过目标模板分支的骨干网络得到特征e_0，e_0通过两条支路分别得到细胞单元c_0，隐藏层状态h_0，每条支路都是由3×3的卷积和1×1的卷积组成。

[0051] $Z = (w+2p) * (h+2p)$ (8)

[0052] $p = (w+h)/4$ (9)

[0053] 步骤2.2：将c_0，h_0和e_0输入到DWConv-LSTM记忆单元，获取初始的目标模板特征F。

[0054] 步骤3：获取跟踪目标在视频当前帧中的对应位置；

[0055] 步骤3.1：从视频当前帧中裁剪出搜索图像，获取多尺度搜索图像对应的特征；

[0056] 假设当前帧是视频的第t帧(t＝1,2,3,…,n)，按照公式(10)，计算出要裁剪的区域面积$A = \{A^{-1}, A^0, A^1\}$，其中，p′表示扩充长度，按照公式(11)计算得到。w和h分别表示给定的跟踪目标的宽和高，k＝1.05，k^i表示k的i次方。然后以算法在视频的上一帧中检测出的目标中心作为中心点，在视频的当前帧中裁剪出边长分别为$\sqrt{A^{-1}}, \sqrt{A^0}, \sqrt{A^1}$的三个正方形区域得到搜索图像，结果如图6所示。再将这些图像尺寸调整成255×255大小，得到多尺度的搜索图像集合$S = \{S_t^{-1}, S_t^0, S_t^1\}$。集合S中的多尺度搜索图像作为一个批次经过搜索图像分支得到多尺度搜索图像对应的特征集合$s = \{s_t^{-1}, s_t^0, s_t^1\}$；

[0057] $A^i = k^i(w+4p') * k^i(h+4p')$，i＝-1,0,1 (10)

[0058] $p' = (w+h)/4$ (11)

[0059] 步骤3.2：获取相似性得分图；

[0060] 当前时刻的目标模板特征F与步骤3.1中得到的多尺度搜索图像对应的特征

$\{s_t^{-1}, s_t^0, s_t^1\}$按照公式(1)进行交叉相关运算，得到相似性得分图集合$r = \{r_t^{-1}, r_t^0, r_t^1\}$；

[0061] 步骤3.3：根据相似性得分图获取跟踪目标对应位置；

[0062] 对集合r中的每一个相似性得分图进行上采样，得到上采样后的相似性得分图集合$R = \{R_t^{-1}, R_t^0, R_t^1\}$，在集合R中找出最大值所在的上采样后的相似性得分图，将其标记为R_t，对R_t中所有值进行比较，获取值最大的K个响应值点，这K个响应值点求平均值得到响应值点d，在视频当前帧中找到d点对应位置即为所寻找的目标所在的位置；

[0063] 步骤4：根据步骤3中找出的当前帧中跟踪目标所在位置裁剪出目标所在区域作为目标模板，将该目标模板输入跟踪模型的目标模板分支获取新的目标模板特征F，读取视频的下一帧作为当前帧，转入步骤3进行下一轮迭代，直至读取完视频中的所有帧结束迭代。

[0064] 为检测本发明的有效性，这里挑选OTB100数据集中的Bird2视频序列进行测试，结果如图7所示，从示例中可以看出本发明能有效应对视频跟踪过程中的遮挡和形态变化等问题，具有很好的鲁棒性。这里又挑选一现实中的场景对本发明方法进行测试，采用与上述实施步骤相同的方法来跟踪目标，结果如图8所示，从示例中可以看出本发明能够很好的跟踪目标，具有一定的使用价值。

[0065] 综上所述，基于记忆单元的孪生网络视觉跟踪方法在学习目标在时间序列上的外观变化趋势的同时又保证了目标在空间上稳定性，能够保证跟踪过程的实时性，提升了跟踪模型面对复杂环境时的跟踪鲁棒性。

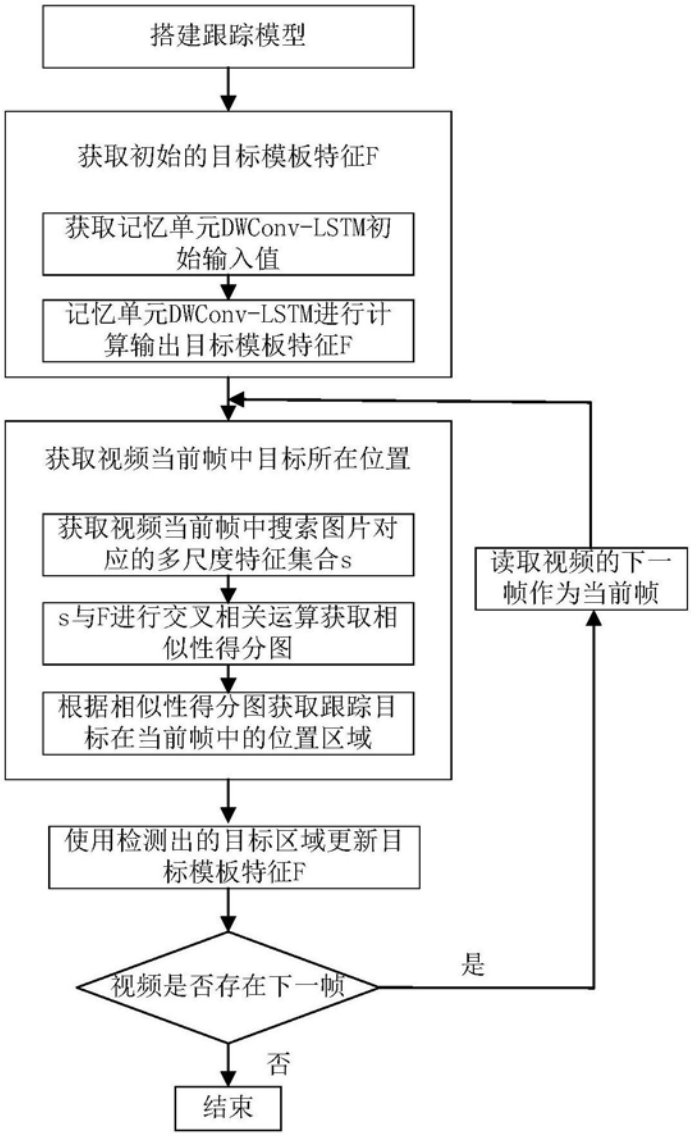

搭建跟踪模型
获取初始的目标模板特征F
获取记忆单元DWConv-LSTM初始输入值
记忆单元DWConv-LSTM进行计算输出目标模板特征F
获取视频当前帧中目标所在位置
获取视频当前帧中搜索图片对应的多尺度特征集合s
s与F进行交叉相关运算获取相似性得分图
根据相似性得分图获取跟踪目标在当前帧中的位置区域
读取视频的下一帧作为当前帧
使用检测出的目标区域更新目标模板特征F
视频是否存在下一帧
是
否
结束

图1

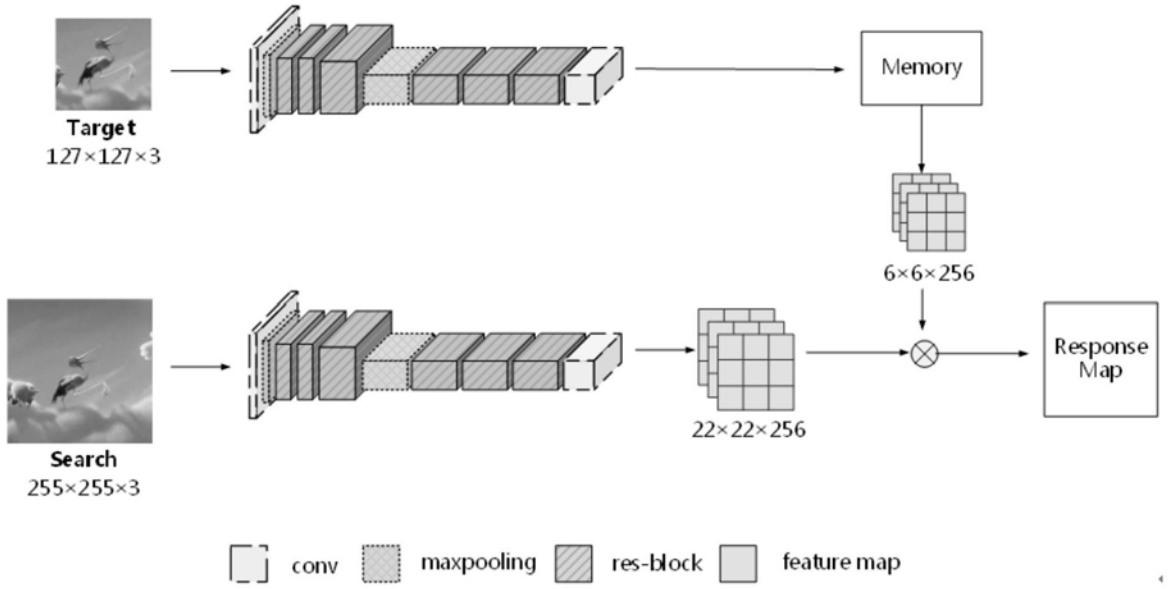

Target
127×127×3
Memory
6×6×256
Search
255×255×3
22×22×256
Response Map
conv
maxpooling
res-block
feature map

图2

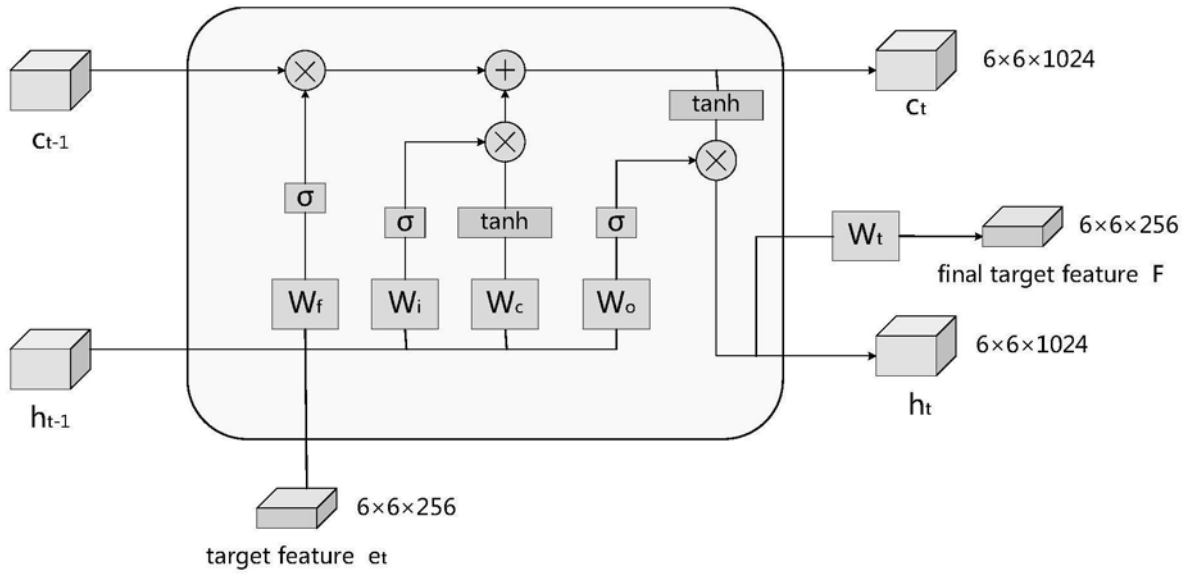

Ct-1
ht-1
σ
σ
tanh
σ
tanh
Wf
Wi
Wc
Wo
Ct
6×6×1024
Wt
6×6×256
final target feature F
ht
6×6×1024
6×6×256
target feature et

图3

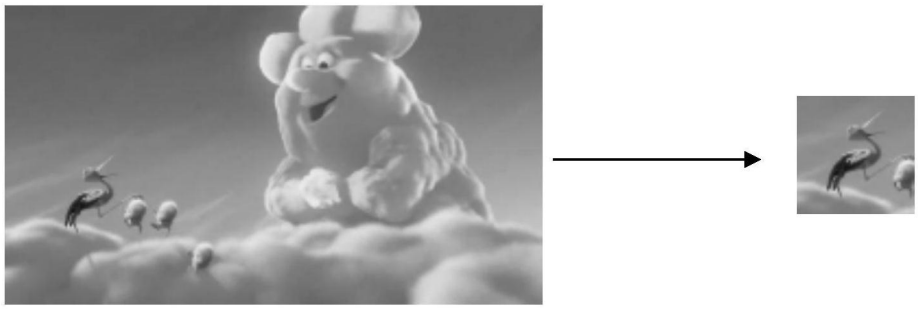


图4

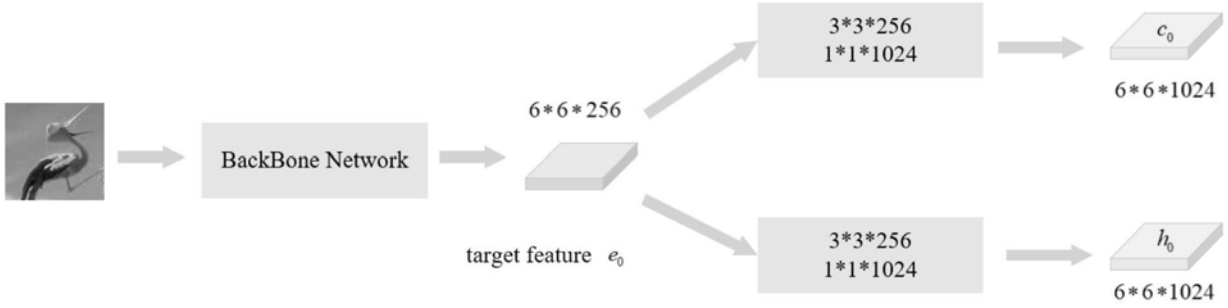

3*3*256
1*1*1024
c_0
6*6*1024
6*6*256
BackBone Network
target feature e_0
3*3*256
1*1*1024
h_0
6*6*1024

图5

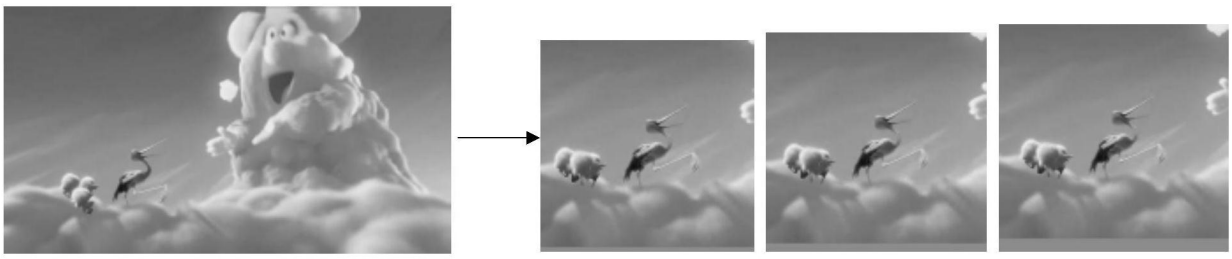


图6

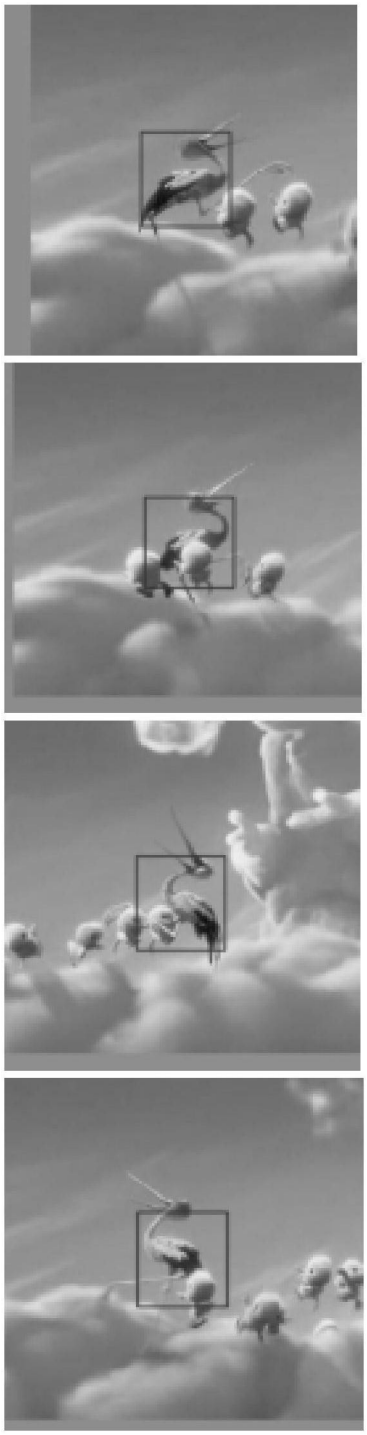


图7

图8