(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 111696137 A
(43)申请公布日 2020.09.22

(21)申请号 202010518472.1

(22)申请日 2020.06.09

(71)申请人 电子科技大学
地址 611731 四川省成都市高新区(西区)西源大道2006号

(72)发明人 王正宁 曾浩 潘力立 何庆东 刘怡君 曾仪 彭大伟

(74)专利代理机构 电子科技大学专利中心 51203
代理人 周刘英

(51)Int.Cl.
G06T 7/246(2017.01)
G06K 9/62(2006.01)

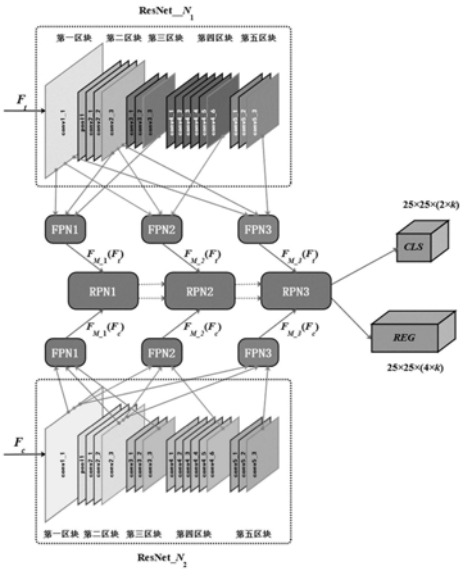
权利要求书3页 说明书7页 附图5页

(54)发明名称

一种基于多层特征混合与注意力机制的目标跟踪方法

(57)摘要

本发明公开了一种基于多层特征混合与注意力机制的目标跟踪方法，该方法利用改进的FPN结构将图像浅层特征加以更好的保留利用，这种对浅层特征有更好保留的改进的FPN结构可以输出具有多维度、多尺度特征的融合特征。对不同尺度大小的目标，以及大小在不断变化的目标拥有更好的跟踪能力。将FPN用于级联的RPN上，对于特征提取更加精准，对于保证跟踪时对于相似干扰物有更好的区分，减少错误跟踪的情况发生。同时，利用注意力机制，在空间尺度上，使得网络对目标可能出现的位置给予更多关注，以减少由目标半遮挡，形变，光照等造成的目标丢失或目标跟踪错误的情况。



CN 111696137 A

1.一种基于多层特征混合与注意力机制的目标跟踪方法,其特征在于,该方法包括以下步骤:

(1)在训练前,对数据集做预处理:训练数据是由视频序列所组成,并带有目标物体位置与大小的标签;目标跟踪网络需要输入的是对应跟踪目标的模板帧和用于寻找目标的搜索帧。将原始视频序列进行裁切处理,获得$w_t \times h_t$像素的模板帧F_t和$w_c \times h_c$像素的搜索帧F_c,,其中模板帧对应视频序列的第一帧,搜索帧对应视频序列的第二帧开始的剩余视频序列;

(2)设计两个并行的5区块深度残差网络N_1、N_2用于提取模板帧和搜索帧的特征,通过权值共享的方式构成孪生网络N_S,使用的深度残差网络将现有的"ResNet-50"的第一个7×7卷积去掉padding,同时将该"ResNet-50"中最后两次步长为2的卷积改成了步长为1的卷积,将模板帧F_t和搜索帧F_c分别送入N_1、N_2,通过卷积、池化、激活等操作,提取出其各自在不同深度的特征;ConvM_N(F_t)和ConvM_N(F_c)分别代表了网络不同层次上模板帧F_t和搜索帧F_c的特征输出,其中M代表该特征图所在的ResNet网络中的区块位置,N代表在某一区块中的具体位置;

(3)设计特征金字塔网络FPN,包括三个FPN:FPN1,FPN2和FPN3分别将从网络N_1、N_2提取的:(Conv1_1、Conv2_3、Conv3_3);(Conv1_1、Conv2_3、Conv4_6);(Conv1_1、Conv2_3、Conv5_3)这3组不同深度的输出特征分别进行融合,获得了3组经过融合的特征,每个FPN接收3个不同尺度的特征图,从大到小、从浅到深分别为F_1、F_2、F_3;特征的融合通过点对点相加完成,通过使用1×1卷积调整一个特征的通道数,使得两个特征通道数相同,再使用2倍上采样或者步长为2的3×3卷积调整另外一个特征的尺寸,使得调整后的两个特征尺寸相同,得以完成点对点相加,即特征融合;将这3种特征进行融合,最终输出融合后的特征F_M,且F_M的尺寸和F_3相同;最终,三个FPN分别输出了模版帧的混合特征$F_{M_1}(F_t)$、$F_{M\_2}(F_t)$、$F_{M_3}(F_t)$和搜索帧的混合特征$F_{M_1}(F_c)$、$F_{M\_2}(F_c)$、$F_{M\_3}(F_c)$;

(4)设计区域推荐网络RPN,包括三个RPN:RPN1,RPN2和RPN3通过分别输入三对模版帧与搜索帧的混合特征:$F_{M\_1}(F_t)$、$F_{M\_1}(F_c)$;$F_{M\_2}(F_t)$、$F_{M\_2}(F_c)$;$F_{M\_3}(F_t)$、$F_{M\_3}(F_c)$,获得建议框的分类结果CLS与回归结果REG;

(5)RPN输出建议框的分类CLS与REG回归结果,这两个不同的输出由两条路径来完成,RPN上半部分输出建议框的分类CLS,下半部分输出建议框的回归REG;RPN首先将从模板帧获取的混合特征$F_M(F_t)$从边缘进行裁切,其中c为当前混合特征通道数,不同组合的混合特征通道数不同;之后通过卷积的调整,将$F_M(F_t)$与$F_M(F_c)$调整到合适的尺寸$[F_M(F_t)]_c$,$[F_M(F_c)]_c$,$[F_M(F_t)]_r$,$[F_M(F_c)]_r$;将$[F_M(F_t)]_c$,$[F_M(F_c)]_c$进行互相关运算得到初步的分类结果CLS_0;将$[F_M(F_t)]_r$,$[F_M(F_c)]_r$进行互相关运算得到初步的回归结果REG_0;

CLS_0的尺寸为$w_{res} \times h_{res} \times 2k$,REG_0的尺寸为$w_{res} \times h_{res} \times 4k$,输出的结果中在$w_{res} \times h_{res}$维度与原图$w_c \times h_c$在空间上呈线性的对应关系,在$w_{res} \times h_{res}$的每一个位置上对应k个预先设定好大小的锚框,锚框的中心为当前所在位置的中心;CLS_0的2k个通道代表了网络预测的k种锚框包含目标的概率P_{pos}和不包含目标的概率P_{neg};REG_0的4k个通道代表了网络预测的k种锚框与实际目标框的长宽差异和位置差异,分别为dx,dy,dw,dh。其与实际目标框的关系为:

$$\begin{cases} dx = \frac{T_x - A_x}{A_w} \\ dy = \frac{T_y - A_y}{A_h} \\ dw = \ln\frac{T_w}{A_w} \\ dh = \ln\frac{T_h}{A_h} \end{cases} \quad (1)$$

其中A_x、A_y表示参考框的中心点，A_w、A_h表示参考框的宽高，T_x、T_y、T_w、T_h表示真值的坐标与长宽，最后通过极大值抑制等方法找出最终的目标；

(6)在输出得到CLS_0和REG_0后，再将其输入空间注意力模块，通过平均池化和最大值池化、卷积、Sigmoid激活操作，获得了$w_{res}\times h_{res}\times 1$的空间注意力权重SA_c和SA_r；CLS_0和REG_0分别与SA_c和SA_r对应位置相乘，并与原始的CLS_0和REG_0相加，或得了最终的RPN输出结果CLS和REG；

(7)对三个RPN：RPN1，RPN2和RPN3的输出结果进行加权相加，作为最终的目标跟踪网络输出结果：

$$\begin{cases} CLS_{all} = \alpha_1 \cdot CLS_1 + \alpha_2 \cdot CLS_2 + \alpha_3 \cdot CLS_3 \\ REG_{all} = \beta_1 \cdot REG_1 + \beta_2 \cdot REG_2 + \beta_3 \cdot REG_3 \end{cases} \quad (2)$$

其中，α_1，α_2，α_3，β_1，β_2，β_3为预先设定的权重。

(8)训练所述目标跟踪网络时的分类损失L_{cls}使用交叉熵损失，回归损失L_{reg}使用具有标准化坐标的平滑L1损失；y表示标签值，$\bar{y}$表示实际分类值，即P_{pos}；dx_T，dy_T，dw_T，dh_T，代表实际k种锚框与实际目标框的长宽差异和位置差异，即真值；损失函数分别定义为：

$$\begin{cases} L_{cls} = -[y\log\bar{y} + (1-y)\log(1-\bar{y})] \\ L_{reg} = smooth_{L1}(dx, dx_T) + smooth_{L1}(dy, dy_T) \\ \quad + smooth_{L1}(dw, dw_T) + smooth_{L1}(dh, dh_T) \end{cases} \quad (3)$$

其中：

$$smooth_{L1}(x,\sigma) = \begin{cases} 0.5\sigma^2 x^2, |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2}, |x| \geq \frac{1}{\sigma^2} \end{cases} \quad (4)$$

最后的损失函数如下：

$loss = L_{cls} + \lambda L_{reg}$ (5)

其中λ是超参数，用于平衡两类损失。

2.根据权利要求1所述的基于多层特征混合与注意力机制的目标跟踪方法，其特征在于，所述步骤(8)训练所述目标跟踪网络具体包括：

对数据集中的视频序列进行处理，根据标签信息，裁切获得127×127像素的模板帧F_t和255×255像素的搜索帧F_c；

将模板帧F_t和搜索帧F_c送入特征提取网络ResNet_N_1与ResNet_N_2，提取出五个不同深度层次的特征，其中两个特征提取网络共享权重；

三个特征金字塔网络，FPN1、FPN2、FPN3分别将提取出的不同深度层次的模板帧F_t与搜索帧F_c特征进行特征融合，其中FPN1融合第一、二、三区块，即一、二、三层获得的特征，FPN2融合第一、二、四区块，即一、二、四层获得的特征，FPN3融合第一、二、五区块，即一、二、五层

获得的特征，三对FPN分别输出了模版帧的混合特征$F_{M_1}(F_t)$、$F_{M\_2}(F_t)$、$F_{M_3}(F_t)$和搜索帧的混合特征$F_{M_1}(F_c)$、$F_{M\_2}(F_c)$、$F_{M\_3}(F_c)$；模版帧的混合特征尺寸都为15×15×512，搜索帧的混合特征尺寸都为31×31×512；

将三对混合特征$F_{M_1}(F_t)$与$F_{M_1}(F_c)$、$F_{M_2}(F_t)$与$F_{M_2}(F_c)$、$F_{M_3}(F_t)$与$F_{M_3}(F_c)$分别送入三个区域推荐网络RPN1，RPN2，RPN3，其中每个区域推荐网络的结构相同，共设置5种锚框，即k＝5；首先将模版帧的混合特征$F_M(F_t)$进行裁切，裁切掉周围部分元素，尺寸变为7×7×512，之后通过四个卷积层调整$F_M(F_t)$与搜索帧的混合特征$F_M(F_c)$的通道数，分别获得：$[F_M(F_t)]_c$，尺寸为5×5×(10×512)；$[F_M(F_t)]_r$，尺寸为5×5×(20×512)；$[F_M(F_c)]_c$，尺寸为29×29×512；$[F_M(F_c)]_r$，尺寸为29×29×512；

分别将$[F_M(F_t)]_c$与$[F_M(F_c)]_c$、$[F_M(F_t)]_r$与$[F_M(F_c)]_r$进行互相关运算，获得分类中间结果CLS_0和回归中间结果REG_0，其中CLS_0的尺寸为25×25×10，REG_0的尺寸为25×25×20。

CLS_0和REG_0分别送入对应的空间注意力模块，获得空间注意力权重SA_c和SA_r；将CLS_0和REG_0与SA_c和SA_r对应位置相乘，并与原始的CLS_0和REG_0相加，获得最终RPN输出分类结果CLS和回归结果REG；CLS和CLS_0尺寸相同；REG和REG_0尺寸相同，“空间注意力”即完成上述步骤；

将RPN1，RPN2，RPN3的输出的分类结果与回归结果按照0.2，0.3，0.5的权值进行加权相加，即获得最终的目标分类结果与建议框回归结果，根据所述式(3)(4)(5)计算损失并进行优化；当达到了设定的训练轮数50轮后，即结束训练进行测试。

一种基于多层特征混合与注意力机制的目标跟踪方法

技术领域

[0001] 本发明属于图像处理和计算机视觉领域，具体涉及一种基于多层特征混合与注意力机制的目标跟踪方法。

背景技术

[0002] 视觉目标跟踪是一项重要的计算机视觉任务，可应用于视觉监控、人机交互、视频压缩等领域。尽管对这一课题进行了广泛的研究，但由于光照变化、部分遮挡、形状变形和相机运动等因素的影响，它在处理复杂的物体外观变化方面仍然存在困难。

[0003] 目标跟踪算法在现阶段主要有两个大的分支，一个是基于相关滤波算法，一个是基于深度学习算法。本发明所提出的目标跟踪方法属于深度学习这一分支。

[0004] 深度学习主要有以下几种方法：卷积神经网络；循环神经网络；生成对抗网络；孪生神经网络。“Learning spatial-aware regressions for visual tracking,C.Sun,D.Wang,H.Lu,a nd M.Yang,in Proc.IEEE CVPR,2018,pp.8962-8970”提出的基于卷积神经网络的目标跟踪方法，构建多个目标模型以捕获各种目标外观，学习不同的目标模型，基于零件的模型、来处理部分遮挡和变形，同时利用双流网络防止过度拟合和学习目标的旋转信息。尽管这一方法在目标估计精度方面取得了大的进展，但这一类基于卷积神经网络的方法仍然具有较高的计算复杂性。发明专利“基于LSTM网络的多机动目标跟踪方法，CN110780290A”基于循环神经网络进行目标跟踪，利用上下文信息来处理相似背景对跟踪目标的影响。但是由于视觉目标跟踪与视频帧的空间和时间信息相关，因此使用基于循环神经网络的方法同时考虑目标的运动。由于模型存在大量的参数导致训练困难，基于循环神经网络的方法数量有限。几乎所有这些方法都试图利用其他信息和内存来改进目标建模。此外，使用基于循环神经网络的方法的第二个目标是避免对预先训练的CNN模型进行微调，这需要大量的时间，而且容易过度拟合。“VITAL:Visual tracking via adversarial learning,Y.Song,C.Ma,X.Wu,L.G ong,L.Bao,W.Zuo,C.Shen,R.W.Lau,and M.H.Yang,in Proc.IEEE CVPR,2018,pp.8990-8999”基于生成对抗网络进行目标跟踪，可以生成所需要的样本，以解决训练样本的不平衡分布问题，同时通过生成样本，解决样本量不足的问题。但生成对抗网络通常很难训练和评估，在实际中对这一问题的解决的技巧性非常强。发明专利“基于卷积神经网络的红外弱小目标检测跟踪方法，CN110728697A”利用孪生网络进行目标跟踪，通过提取图片的深度特征，进行特征的匹配进而完成目标的跟踪，但该方法对目标浅层特征利用不足，同时对跟踪中的遮挡、半遮挡、光照变化、形变等问题没有好的解决，方法的鲁棒性有待提升。

[0005] 针对以往深度学习目标特征利用不均以及被跟踪物体受到的遮挡、半遮挡、光照变化、形变等问题，本发明以孪生网络为基础，利用多个FPN进行浅层与深层特征的结合，同时使用注意力机制，提高方法的鲁棒性。

发明内容

[0006] 本发明属于计算机视觉和深度学习领域，通过改进孪生网络的特征提取部分和区域推荐网络部分，使得整个目标跟踪网络拥有更强的特征提取能力和更强的鲁棒性。本发明提出的一种基于多层特征混合与注意力机制的目标跟踪方法具体步骤如下：

[0007] (1)在训练前，对数据集做预处理：训练数据是由视频序列所组成，并带有目标物体

[0008] 位置与大小的标签；目标跟踪网络需要输入的是对应跟踪目标的模板帧和用于寻找目标的搜索帧。将原始视频序列进行裁切处理，获得$w_t \times h_t$像素的模板帧F_t和$w_c \times h_c$像素的搜索帧F_c，，其中模板帧对应视频序列的第一帧，搜索帧对应视频序列的第二帧开始的剩余视频序列。

[0009] (2)设计两个并行的5区块深度残差网络N_1、N_2用于提取模板帧和搜索帧的特征，通过权值共享的方式构成孪生网络N_S，使用的深度残差网络将现有的“ResNet-50”的第一个7×7卷积去掉padding，同时将该“ResNet-50”中最后两次步长为2的卷积改成了步长为1的卷积，将模板帧F_t和搜索帧F_c分别送入N_1、N_2，通过卷积、池化、激活等操作，提取出其各自在不同深度的特征；ConvM_N(F_t)和ConvM_N(F_c)分别代表了网络不同层次上模板帧F_t和搜索帧F_c的特征输出，其中M代表该特征图所在的ResNet网络中的区块位置，N代表在某一区块中的具体位置。

[0010] (3)设计特征金字塔网络FPN，包括三个FPN：FPN1，FPN2和FPN3分别将从网络N_1、N_2提取的：(Conv1_1、Conv2_3、Conv3_3)；(Conv1_1、Conv2_3、Conv4_6)；(Conv1_1、Conv2_3、Conv5_3)这3组不同深度的输出特征分别进行融合，获得了3组经过融合的特征，每个FPN接收3个不同尺度的特征图，从大到小、从浅到深分别为F_1、F_2、F_3；特征的融合通过点对点相加完成，通过使用1×1卷积调整一个特征的通道数，使得两个特征通道数相同，再使用2倍上采样或者步长为2的3×3卷积调整另外一个特征的尺寸，使得调整后的两个特征尺寸相同，得以完成点对点相加，即特征融合；将这3种特征进行融合，最终输出融合后的特征F_M，且F_M的尺寸和F_3相同；最终，三个FPN分别输出了模版帧的混合特征$F_{M_1}(F_t)$、$F_{M\_2}(F_t)$、$F_{M_3}(F_t)$和搜索帧的混合特征$F_{M_1}(F_c)$、$F_{M\_2}(F_c)$、$F_{M\_3}(F_c)$；

[0011] (4)设计区域推荐网络RPN，包括三个RPN：RPN1，RPN2和RPN3通过分别输入三对模版帧与搜索帧的混合特征：$F_{M\_1}(F_t)$、$F_{M\_1}(F_c)$；$F_{M\_2}(F_t)$、$F_{M\_2}(F_c)$；$F_{M\_3}(F_t)$、$F_{M\_3}(F_c)$，获得建议框的分类结果CLS与回归结果REG；

[0012] (5)RPN输出建议框的分类CLS与REG回归结果，这两个不同的输出由两条路径来完成，RPN上半部分输出建议框的分类CLS，下半部分输出建议框的回归REG；RPN首先将从模板帧获取的混合特征$F_M(F_t)$从边缘进行裁切，其中c为当前混合特征通道数，不同组合的混合特征通道数不同；之后通过卷积的调整，将$F_M(F_t)$与$F_M(F_c)$调整到合适的尺寸$[F_M(F_t)]_c$，$[F_M(F_c)]_c$，$[F_M(F_t)]_r$，$[F_M(F_c)]_r$；将$[F_M(F_t)]_c$，$[F_M(F_c)]_c$进行互相关运算得到初步的分类结果CLS_0；将$[F_M(F_t)]_r$，$[F_M(F_c)]_r$进行互相关运算得到初步的回归结果REG_0；

[0013] CLS_0的尺寸为$w_{res} \times h_{res} \times 2k$，REG_0的尺寸为$w_{res} \times h_{res} \times 4k$，输出的结果中在$w_{res} \times h_{res}$维度与原图$w_c \times h_c$在空间上呈线性的对应关系，在$w_{res} \times h_{res}$的每一个位置上对应k个预先设定好大小的锚框，锚框的中心为当前所在位置的中心；CLS_0的2k个通道代表了网络预测的k种锚框包含目标的概率P_{pos}和不包含目标的概率P_{neg}；REG_0的4k个通道代表了网络

预测的k种锚框与实际目标框的长宽差异和位置差异，分别为dx，dy，dw，dh。其与实际目标框的关系为：

[0014] $$\begin{cases} dx = \frac{T_x - A_x}{A_w} \\ dy = \frac{T_y - A_y}{A_h} \\ dw = \ln \frac{T_w}{A_w} \\ dh = \ln \frac{T_h}{A_h} \end{cases} \quad (1)$$

[0015] 其中A_x、A_y表示参考框的中心点，A_w、A_h表示参考框的宽高，T_x、T_y、T_w、T_h表示真值的坐标与长宽，最后通过极大值抑制等方法找出最终的目标；

[0016] (6)在输出得到CLS_O和REG_O后，再将其输入空间注意力模块，通过平均池化和最大值池化、卷积、Sigmoid激活操作，获得了$w_{res} \times h_{res} \times 1$的空间注意力权重SA_c和SA_r；CLS_O和REG_O分别与SA_c和SA_r对应位置相乘，并与原始的CLS_O和REG_O相加，或得了最终的RPN输出结果CLS和REG；

[0017] (7)对三个RPN：RPN1，RPN2和RPN3的输出结果进行加权相加，作为最终的目标跟踪网络输出结果：

[0018] $$\begin{cases} CLS_{all} = \alpha_1 \cdot CLS_1 + \alpha_2 \cdot CLS_2 + \alpha_3 \cdot CLS_3 \\ REG_{all} = \beta_1 \cdot REG_1 + \beta_2 \cdot REG_2 + \beta_3 \cdot REG_3 \end{cases} \quad (2)$$

[0019] 其中，α_1，α_2，α_3，β_1，β_2，β_3为预先设定的权重。

[0020] (8)训练所述目标跟踪网络时的分类损失L_{cls}使用交叉熵损失，回归损失L_{reg}使用具有标准化坐标的平滑L1损失；y表示标签值，y表示实际分类值，即P_{pos}；dx_T，dy_T，dw_T，dh_T，代表实际k种锚框与实际目标框的长宽差异和位置差异，即真值；损失函数分别定义为：

[0021] $$\begin{cases} L_{cls} = -[y \log \bar{y} + (1-y) \log(1-\bar{y})] \\ L_{reg} = smooth_{L1}(dx, dx_T) + smooth_{L1}(dy, dy_T) \\ \quad + smooth_{L1}(dw, dw_T) + smooth_{L1}(dh, dh_T) \end{cases} \quad (3)$$

[0022] 其中：

[0023] $$smooth_{L1}(x, \sigma) = \begin{cases} 0.5\sigma^2 x^2, |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2}, |x| \geq \frac{1}{\sigma^2} \end{cases} \quad (4)$$

[0024] 最后的损失函数如下：

[0025] $loss = L_{cls} + \lambda L_{reg}$ (5)

[0026] 其中λ是超参数，用于平衡两类损失。

[0027] 本发明利用改进的FPN结构。相较于传统FPN中获得的深层特征对浅层特征保留不足的情况，利用改进的FPN结构，将图像浅层特征加以更好的保留利用。这种对浅层特征有更好保留的改进的FPN结构可以输出具有多维度、多尺度特征的融合特征。对不同尺度大小的目标，以及大小在不断变化的目标拥有更好的跟踪能力。将FPN用于级联的RPN上，对于特征提取更加精准，对于保证跟踪时对于相似干扰物有更好的区分，减少错误跟踪的情况发生。同时，利用注意力机制，在空间尺度上，使得网络对目标可能出现的位置给予更多关注，以减少由目标半遮挡，形变，光照等造成的目标丢失或目标跟踪错误的情况。

附图说明

[0028] 图1为本发明的模板帧与搜索帧示意图

[0029] 图2为本发明的目标跟踪网络整体结构图

[0030] 图3为本发明的FPN结构图

[0031] 图4为本发明的RPN结构图

[0032] 图5为本发明RPN输出结果示意图

[0033] 图6为本发明的目标跟踪网络训练流程图

具体实施方式

[0034] 下面结合附图对本发明的具体实施方式以及工作原理作进一步详细说明。

[0035] 本发明提出的一种基于多层特征混合与注意力机制的目标跟踪方法具体步骤如下：

[0036] (1)在训练前，对数据集做预处理。训练数据是由视频序列所组成，并带有目标物体位置与大小的标签。目标跟踪网络需要输入的是对应跟踪目标的模板帧和用于寻找目标的搜索帧。将原始视频序列进行裁切处理，获得$w_t \times h_t$像素的模板帧F_t和$w_c \times h_c$像素的搜索帧F_c，如图1、图2所示。其中模板帧对应视频序列的第一帧，搜索帧对应视频序列的第二帧开始的剩余视频序列。

[0037] (2)设计两个并行的5区块深度残差网络N_1、N_2用于提取模板帧和搜索帧的特征，通过权值共享的方式构成孪生网络N_S。使用的深度残差网络将现有的“ResNet-50”的第一个7×7卷积去掉padding，同时将该“ResNet-50”中最后两次步长为2的卷积改成了步长为1的卷积。将模板帧F_t和搜索帧F_c分别送入N_1、N_2，通过卷积、池化、激活等操作，提取出其各自在不同深度的特征。ConvM_N(F_t)和ConvM_N(F_c)分别代表了网络不同层次上模板帧F_t和搜索帧F_c的特征输出，其中M代表该特征图所在的ResNet网络中的区块位置，N代表在某一区块中的具体位置。

[0038] (3)设计特征金字塔网络(Feature Pyramid Networks，FPN)，三个FPN(FPN1，FPN2，FPN3)分别将从网络N_1、N_2提取的：(Conv1_1、Conv2_3、Conv3_3)；(Conv1_1、Conv2_3、Conv4_6)；(Conv1_1、Conv2_3、Conv5_3)这3组不同深度的输出特征分别进行融合，获得了3组经过融合的特征。

[0039] 本发明使用的单个FPN的具体结构如图4所示。每个FPN接收3个不同尺度的特征图，从大到小、从浅到深分别为F_1、F_2、F_3。特征的融合通过点对点相加完成，通过使用1×1卷积调整一个特征的通道数，使得两个特征通道数相同，再使用2倍上采样或者步长为2的3×3卷积调整另外一个特征的尺寸，使得调整后的两个特征尺寸相同，得以完成点对点相加，即特征融合。将这3种特征进行融合，最终输出融合后的特征F_M，且F_M的尺寸和F_3相同。最终，三个FPN分别输出了模版帧的混合特征$F_{M_1}(F_t)$、$F_{M\_2}(F_t)$、$F_{M_3}(F_t)$和搜索帧的混合特征$F_{M_1}(F_c)$、$F_{M\_2}(F_c)$、$F_{M\_3}(F_c)$。

[0040] (4)区域推荐网络(Region Proposal Network，RPN)，三个RPN(RPN1，RPN2，RPN3)通过分别输入三对模版帧与搜索帧的混合特征：$F_{M\_1}(F_t)$、$F_{M\_1}(F_c)$；$F_{M\_2}(F_t)$、$F_{M\_2}(F_c)$；$F_{M\_3}(F_t)$、$F_{M\_3}(F_c)$，获得建议框的分类结果CLS与回归结果REG，如图2所示。

[0041] (5)RPN需要输出建议框的分类CLS与REG回归结果，这两个不同的输出需要两条路

径来完成，图2中的RPN上半部分输出建议框的分类CLS，下半部分输出建议框的回归REG。RPN首先将从模板帧获取的混合特征$F_M(F_t)$从边缘进行裁切，其中c为当前混合特征通道数，不同组合的混合特征通道数不同。之后通过卷积的调整，将$F_M(F_t)$与$F_M(F_c)$调整到合适的尺寸$[F_M(F_t)]_c$，$[F_M(F_c)]_c$，$[F_M(F_t)]_r$，$[F_M(F_c)]_r$。将$[F_M(F_t)]_c$，$[F_M(F_c)]_c$进行互相关运算得到初步的分类结果CLS_0；将$[F_M(F_t)]_r$，$[F_M(F_c)]_r$进行互相关运算得到初步的回归结果REG_0。

[0042] CLS_0的尺寸为$w_{res}\times h_{res}\times 2k$，REG_0的尺寸为$w_{res}\times h_{res}\times 4k$，如图5所示，输出的结果中在$w_{res}\times h_{res}$维度与原图$w_c\times h_c$在空间上呈线性的对应关系，在$w_{res}\times h_{res}$的每一个位置上对应k个预先设定好大小的锚框，锚框的中心为当前所在位置的中心。CLS_0的2k个通道代表了网络预测的k种锚框包含目标的概率P_{pos}和不包含目标的概率P_{neg}。REG_0的4k个通道代表了网络预测的k种锚框与实际目标框的长宽差异和位置差异，分别为dx，dy，dw，dh。其与实际目标框的关系为：

[0043]
$$\begin{cases} dx = \frac{T_x - A_x}{A_w} \\ dy = \frac{T_y - A_y}{A_h} \\ dw = \ln\frac{T_w}{A_w} \\ dh = \ln\frac{T_h}{A_h} \end{cases} \tag{1}$$

[0044] 其中A_x、A_y表示参考框的中心点，A_w、A_h表示参考框的宽高，T_x、T_y、T_w、T_h表示真值的坐标与长宽。最后通过极大值抑制等方法找出最终的目标。

[0045] (6)在输出得到CLS_0和REG_0后，再将其输入空间注意力模块，如图4所示，通过平均池化和最大值池化、卷积、Sigmoid激活操作，获得了$w_{res}\times h_{res}\times 1$的空间注意力权重SA_c和SA_r。CLS_0和REG_0分别与SA_c和SA_r对应位置相乘，并与原始的CLS_0和REG_0相加，或得了最终的RPN输出结果CLS和REG。

[0046] (7)对三个RPN(RPN1，RPN2，RPN3)的输出结果进行加权相加，作为最终的目标跟踪网络输出结果：

[0047]
$$\begin{cases} CLS_{all} = \alpha_1 \cdot CLS_1 + \alpha_2 \cdot CLS_2 + \alpha_3 \cdot CLS_3 \\ REG_{all} = \beta_1 \cdot REG_1 + \beta_2 \cdot REG_2 + \beta_3 \cdot REG_3 \end{cases} \tag{2}$$

[0048] 其中，α_1，α_2，α_3，β_1，β_2，β_3为预先设定的权重。

[0049] (8)训练所述目标跟踪网络时的分类损失L_{cls}使用交叉熵损失，回归损失L_{reg}使用具有标准化坐标的平滑L1损失。y表示标签值，y表示实际分类值(即P_{pos})；dx_T，dy_T，dw_T，dh_T，代表实际k种锚框与实际目标框的长宽差异和位置差异，即真值。损失函数分别定义为：

[0050]
$$\begin{cases} L_{cls} = -[y\log\bar{y} + (1-y)\log(1-\bar{y})] \\ L_{reg} = smooth_{L1}(dx, dx_T) + smooth_{L1}(dy, dy_T) \\ \quad + smooth_{L1}(dw, dw_T) + smooth_{L1}(dh, dh_T) \end{cases} \tag{3}$$

[0051] 其中：

[0052]
$$smooth_{L1}(x,\sigma) = \begin{cases} 0.5\sigma^2x^2, |x| < \frac{1}{\sigma^2} \\ |x| - \frac{1}{2\sigma^2}, |x| \geq \frac{1}{\sigma^2} \end{cases} \tag{4}$$

[0053] 最后的损失函数如下：

[0054] $loss=L_{cls}+\lambda L_{reg}$ (5)

[0055] 其中λ是超参数，用于平衡两类损失。

[0056] 本发明的一种实施例所涉及的关键参数如表1所示，附录1部分图中标注的具体参数是以该实施参数为基准：

[0057] 表1一种实施例参数

[0058]

	设计参数	实施参数
模板帧 F_t	$w_t \times h_t$	127×127
搜索帧 F_c	$w_c \times h_c$	255×255
分类结果 *CLS/ CLS_O*	$w_{res} \times h_{res} \times 2k$	25×25×10
回归结果 *REG/ CLS_O*	$w_{res} \times h_{res} \times 4k$	25×25×20

[0059] 本发明所设计的目标跟踪网络具体训练流程如图6所示，其中具体训练过程以及该方案具体实施相关参数如下：

[0060] 对数据集中的视频序列进行处理。根据标签信息，裁切获得127×127像素的模板帧F_t和255×255像素的搜索帧F_c。

[0061] 将模板帧F_t和搜索帧F_c送入图2中的特征提取网络ResNet_N_1与ResNet_N_2，提取出五个不同深度层次的特征，其中两个特征提取网络共享权重。

[0062] 三个特征金字塔网络，如图3所示，FPN1、FPN2、FPN3分别将提取出的不同深度层次的模板帧F_t与搜索帧F_c特征进行特征融合，其中FPN1融合第一、二、三区块(层)获得的特征，FPN2融合第一、二、四区块(层)获得的特征，FPN3融合第一、二、五区块(层)获得的特征，如图2所示。三对FPN分别输出了模版帧的混合特征$F_{M_1}(F_t)$、$F_{M\_2}(F_t)$、$F_{M_3}(F_t)$和搜索帧的混合特征$F_{M_1}(F_c)$、$F_{M\_2}(F_c)$、$F_{M\_3}(F_c)$。模版帧的混合特征尺寸都为15×15×512，搜索帧的混合特征尺寸都为31×31×512。

[0063] 将三对混合特征$F_{M_1}(F_t)$与$F_{M_1}(F_c)$、$F_{M_2}(F_t)$与$F_{M_2}(F_c)$、$F_{M_3}(F_t)$与$F_{M_3}(F_c)$分别送入三个区域推荐网络RPN1，RPN2，RPN3，如图2所示。其中每个区域推荐网络的结构相同，如图4所示，共设置5种锚框，即k=5。首先将模版帧的混合特征$F_M(F_t)$进行裁切，裁切掉周围部分元素，尺寸变为7×7×512，之后通过四个卷积层调整$F_M(F_t)$与搜索帧的混合特征$F_M(F_c)$的通道数，分别可以获得：$[F_M(F_t)]_c$，尺寸为5×5×(10×512)；$[F_M(F_t)]_r$，尺寸为5×5×(20×512)；$[F_M(F_c)]_c$，尺寸为29×29×512；$[F_M(F_c)]_r$，尺寸为29×29×512。

[0064] 分别将$[F_M(F_t)]_c$与$[F_M(F_c)]_c$、$[F_M(F_t)]_r$与$[F_M(F_c)]_r$进行互相关运算，可以获得分类中间结果CLS_O和回归中间结果REG_O。其中CLS_O的尺寸为25×25×10，REG_O的尺寸为25×25×20。

[0065] CLS_O和REG_O分别送入对应的空间注意力模块，获得空间注意力权重SA_c和SA_r。将CLS_O和REG_O与SA_c和SA_r对应位置相乘，并与原始的CLS_O和REG_O相加，获得最终RPN输出分类结果CLS和回归结果REG。CLS和CLS_O尺寸相同；REG和REG_O尺寸相同。流程图中的“空间注意力”即完成上述步骤。

[0066] 将RPN1，RPN2，RPN3的输出的分类结果与回归结果按照0.2，0.3，0.5的权值进行加

权相加,即获得最终的目标分类结果与建议框回归结果。根据式(3)(4)(5)计算损失并进行优化。当达到了设定的训练轮数50轮后,即结束训练进行测试。

[0067] 以上所述,仅为本发明的具体实施方式,本说明书中所公开的任一特征,除非特别叙述,均可被其他等效或具有类似目的的替代特征加以替换;所公开的所有特征、或所有方法或过程中的步骤,除了互相排斥的特征和/或步骤以外,均可以任何方式组合;本领域的技术人员根据本发明技术方案的技术特征所做出的任何非本质的添加、替换,均属于本发明的保护范围。

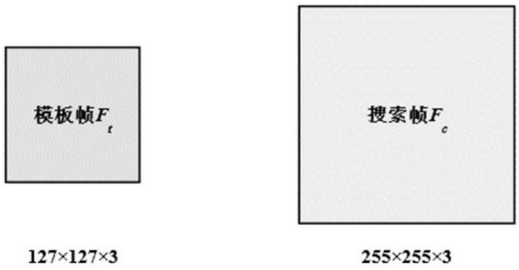

模板帧F_t
搜索帧F_c
127×127×3
255×255×3

图1

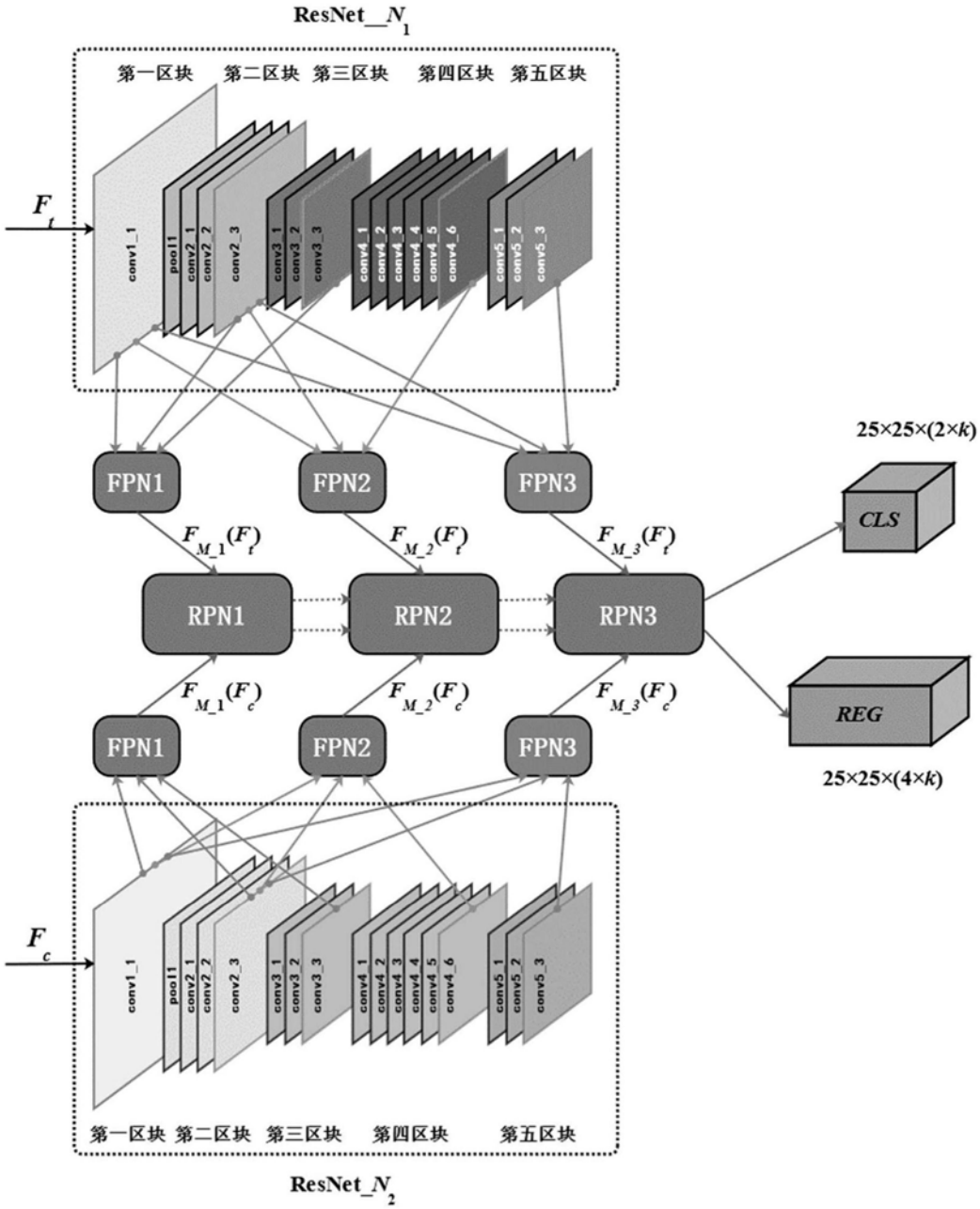

ResNet_N_1
第一区块 第二区块 第三区块 第四区块 第五区块
F_t
conv1_1
pool1
conv2_1
conv2_2
conv2_3
conv3_1
conv3_2
conv3_3
conv4_1
conv4_2
conv4_3
conv4_4
conv4_5
conv4_6
conv5_1
conv5_2
conv5_3
FPN1
FPN2
FPN3
$F_{M\_1}(F_t)$
$F_{M\_2}(F_t)$
$F_{M\_3}(F_t)$
RPN1
RPN2
RPN3
25×25×(2×k)
CLS
REG
25×25×(4×k)
$F_{M\_1}(F_c)$
$F_{M\_2}(F_c)$
$F_{M\_3}(F_c)$
FPN1
FPN2
FPN3
F_c
conv1_1
pool1
conv2_1
conv2_2
conv2_3
conv3_1
conv3_2
conv3_3
conv4_1
conv4_2
conv4_3
conv4_4
conv4_5
conv4_6
conv5_1
conv5_2
conv5_3
第一区块 第二区块 第三区块 第四区块 第五区块
ResNet_N_2

图2

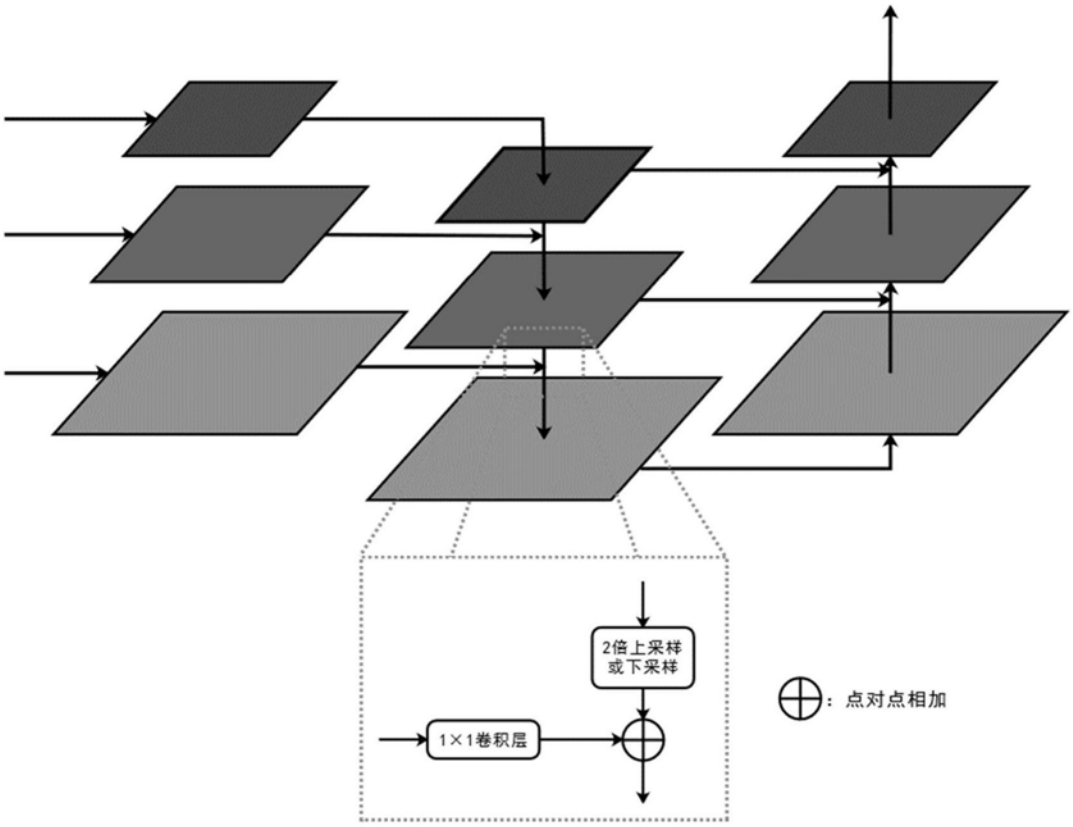


图3

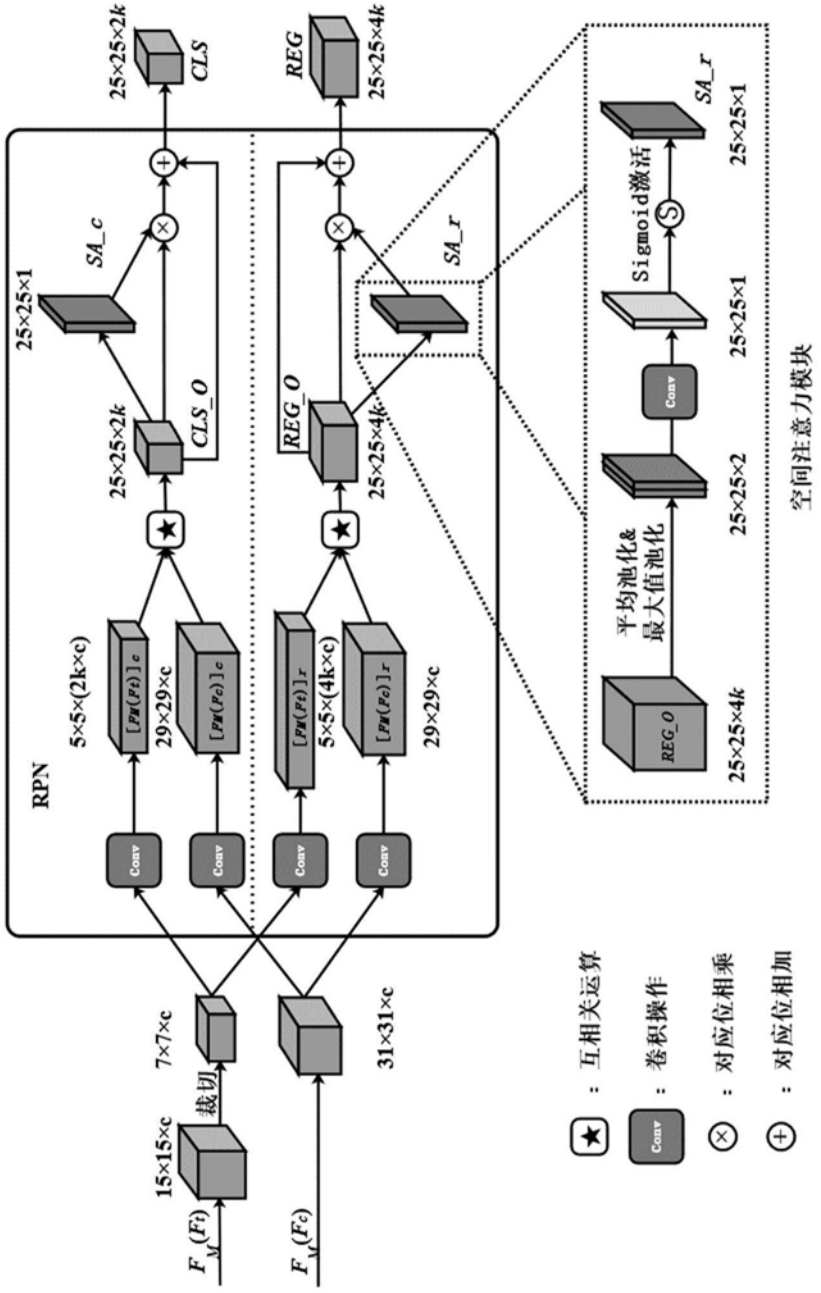

RPN
裁切
7×7×c
15×15×c
31×31×c
5×5×(2k×c)
29×29×c
5×5×(4k×c)
29×29×c
Conv
25×25×2k
CLS
CLS_O
SA_c
25×25×1
25×25×4k
REG
REG_O
SA_r
平均池化&
最大值池化
25×25×2
Sigmoid激活
空间注意力模块
：互相关运算
：卷积操作
：对应位相乘
：对应位相加

图4

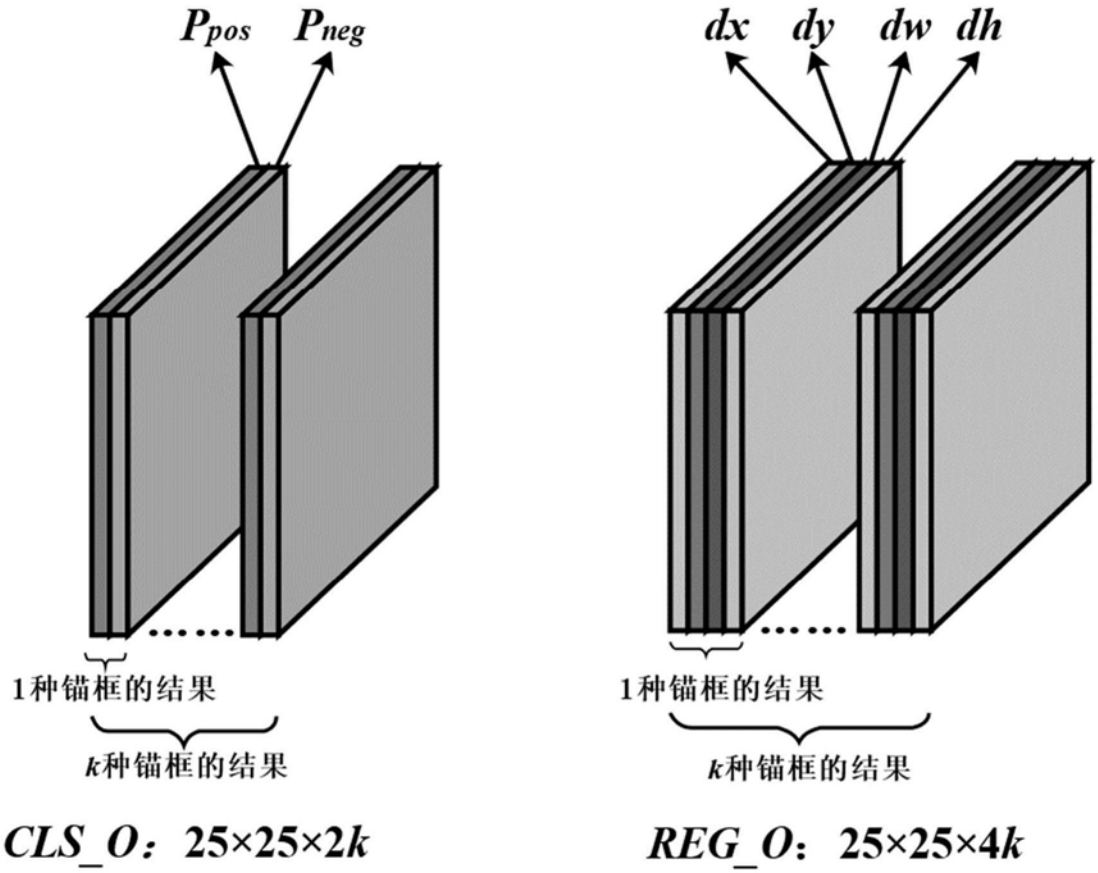

Ppos
Pneg
1种锚框的结果
k种锚框的结果
CLS_O：25×25×2k
dx
dy
dw
dh
1种锚框的结果
k种锚框的结果
REG_O：25×25×4k

图5

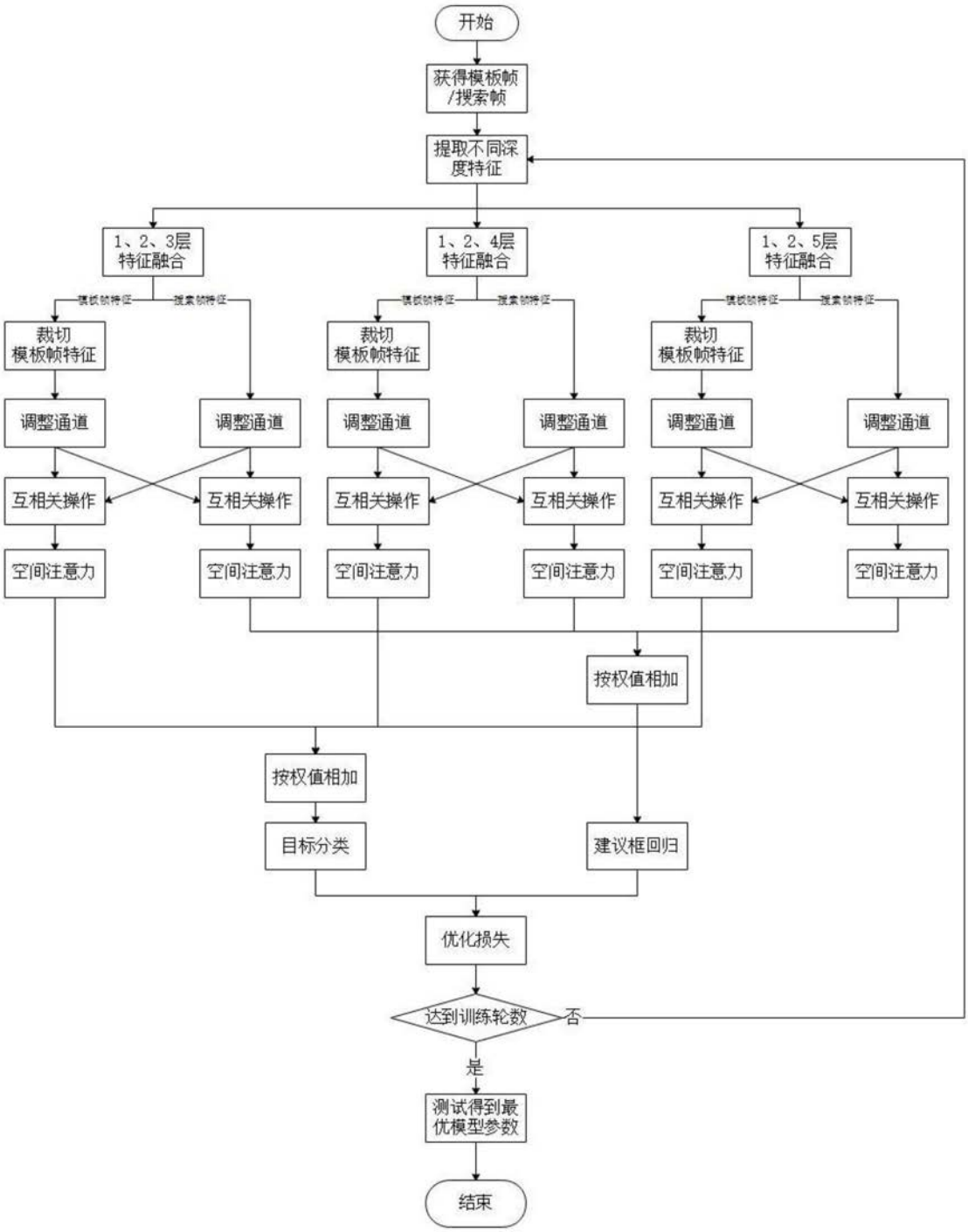

开始
获得模板帧/搜索帧
提取不同深度特征
1、2、3层特征融合
1、2、4层特征融合
1、2、5层特征融合
模板帧特征
搜索帧特征
裁切模板帧特征
调整通道
互相关操作
空间注意力
按权值相加
目标分类
建议框回归
优化损失
达到训练轮数
否
是
测试得到最优模型参数
结束

图6