(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 112949673 B
(45)授权公告日 2023.04.07

(21)申请号 201911270269.0

(22)申请日 2019.12.11

(65)同一申请的已公布的文献号
申请公布号 CN 112949673 A

(43)申请公布日 2021.06.11

(73)专利权人 四川大学
地址 610065 四川省成都市武侯区一环路南一段24号

(72)发明人 吴晓红 吴稳稳 何小海 刘强 陈洪刚 卿粼波 吴小强

(51)Int.Cl.
G06V 10/80(2022.01)
G06V 10/82(2022.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 108510012 A,2018.09.07
CN 110414600 A,2019.11.05
WO 2017080929 A1,2017.05.18
张松 等."一种多特征融合的运动目标检测算法".《扬州大学学报(自然科学版)》.2018,第21卷(第4期),42-46.

审查员 张帆

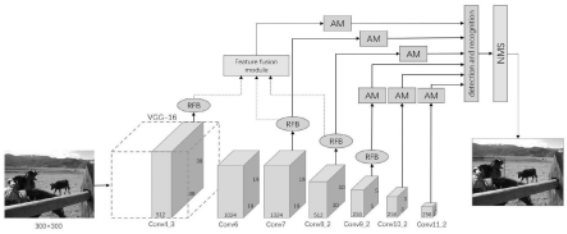
权利要求书2页 说明书5页 附图2页

(54)发明名称

一种基于全局注意力的特征融合目标检测与识别方法

(57)摘要

本发明公开了一种基于全局注意力的特征融合目标检测与识别方法。包括以下步骤:首先由卷积神经网络提取六层不同尺度大小的特征图,然后采用多级特征融合的方法,将浅层和深层特征的语义信息相结合,提高特征图的表达能力。接着引入全局注意力模块来结合上下文信息,增强有效特征和抑制冗余特征。此外,在多任务损失函数的基础上,增加一项额外的惩罚项来平衡正负样本。最后通过训练,不断优化网络参数得到最终的检测模型。本发明所提方法在检测的精度和速度上都有一定的提高,并提升了小目标物体的检测效果,在人机交互、人脸识别、计算摄影、自动驾驶、视频监控等各个方面都有着重要的研究价值和应用前景。



CN 112949673 B

1.一种基于全局注意力的特征融合目标检测与识别方法,其特征在于包括以下步骤:

(1)使用基准网络VGG-16作为特征抽取网络,末端辅以一系列卷积和池化层,再结合空洞卷积模块,初步得到多尺度的卷积特征层;

(2)构建多级特征融合模块,将第1、2、3层特征图进行融合,将深层和浅层的特征语义信息相结合,从而得到更加有效的特征;

(3)构建包含Context Modeling、Transform和Fusion三个部分组成的全局注意力模块,以Context Modeling捕捉上下文特征图像素之间的关系,并以Transform对通道间特征进行建模,自适应地重新标定通道特征响应,最终以Fusion聚合处理后的全局上下文特征到原始特征上,从而得到更加有效且丰富的特征,提高特征图的表达能力;

(4)在多任务损失函数的基础引入了Focal loss来调整正负样本的平衡,然后将上述处理后的特征进行分类和边框回归操作,通过训练模型,不断优化网络参数,最后通过NMS过滤重复检测边框得到最终的检测模型。

2.根据权利要求1所述的方法,其特征在于(1)中获取多尺度的卷积特征层,获取方法如下:

本发明采用VGG16作为基础网络,将VGG16后的两个全连接层FC6和FC7转换成普通的卷积层Conv6和Conv7,之后又添加多个卷积和池化层,然后从后面新增的卷积层中选取Conv7,Conv8_2,Conv9_2,Conv10_2,Conv11_2加上Conv4_3层共6个特征图作为检测所用的特征图,并将Conv4_3、Conv7、Conv8_2和Conv9_2通过空洞卷积模块,初步得到多尺度的卷积特征层。

3.根据权利要求1所述的方法,其特征在于(2)中多级特征融合模块,融合方法如下:

本发明将初步得到的第1、2、3层特征图进行卷积或上采样操作,分别将该3层特征变换到指定大小的尺寸和通道后进行Concate融合,再经过卷积W_k和ReLU操作得到融合后的特征图,经过多级特征融合得到的特征图可以表示为

$$x_1 = ReLU\left\{W_k\left\{Concate\left[\mathcal{T}_i(x_1), \mathcal{T}_i(x_2), \mathcal{T}_i(x_3)\right]\right\}\right\} \quad (1)$$

其中x_i表示第i层特征图,$x_i \in R^{H\times W\times C}$,H、W和C分别表示特征图的长、宽和通道,$\mathcal{T}_i$表示对特征图进行卷积和上采样操作。

4.根据权利要求1所述的方法,其特征在于(3)中构建全局注意力增强模块,构建方法如下:

全局注意力模块分为三个部分,分别是Context Modeling、Transform和Fusion,Context Modeling是上下文建模部分,采用卷积和Softmax操作来获取注意力的权值,将全局上下文建模为所有位置特征的加权平均值,然后聚集全局上下文特征到每个位置的特征上,定义x为输入的特征图,特征图的宽和高分别为W和H,$x = \{x_i\}_{i=1}^{N_P}$,x_i和x_j分别表示某一位置的像素值,x_i经过Context Modeling得到的表达为

$$y_i = \sum_{j=1}^{N_P} \alpha_j x_j \quad (2)$$

式中N_P为特征图的位置数量,$N_P = H \cdot W$,α_j用来计算位置i和所有可能关联的位置j之间

的关系，获取全局上下文信息的权重，$\alpha_j = \frac{\exp(W_k x_j)}{\sum_{m=1}^{N_P} \exp(W_k x_m)}$，i表示该特征图中具体位置的索引，j是所有可能的位置的索引，位置i和j两点之间的相似性关联函数通过$\exp(W_k x_j)$表征，W_k是1×1的卷积操作，这里看做是一个线性转换矩阵；Transform是特征转换部分，它通过卷积、Globalpooling和ReLU操作实现，用来捕获通道间的依赖关系，该过程可以表式为

$s = x_c \cdot \sigma\{W_u \delta\{LN[W_r F_g(x_c)]\}\}$ (3)

其中x_c是对ContextModeling模块的输出进行变换和卷积W_u操作得到的特征，δ是ReLU操作，σ是Sigmoid操作；接着在x_c的每个通道上执行全局平均池化，在空间维度上对特征进行压缩，使其具有全局的感受野，池化的过程F_g表示为

$$F_g(x_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \quad (4)$$

F_g操作将特征空间上所有点的信息平均成了一个值，表征在特征通道上响应的全局分布；接着通过W_r卷积操作减少通道数从而降低计算量，加入LayerNorm作为正则化提高泛化性；然后通过W_u卷积操作恢复到原通道大小，以学习的方式为每个特征通道生成权重，来显式地建模特征通道间的相关性；最后经过Sigmoid激活函数层，获得0～1之间归一化的权重，通过乘法逐通道将归一化后的权重加权到原来的特征上，完成在通道维度上的对原始特征的重标定；Fusion是特征融合部分，用于将处理得到的全局上下文特征聚合到原始特征上，得到最终输出特征的表达式为

$z = x + s$ (5)。

5.根据权利要求1所述的方法，其特征在于(4)中损失函数的优化，优化方法如下：

在多任务损失函数的基础上，加入了Focal loss惩罚项，对损失函数进行了改进，改善了单级探测器所面临的类别不平衡的问题，总目标损失函数由每个默认框的定位损失与分类损失的加权和构成，即

$$L(x,c,l,g) = \frac{1}{N}(L_{loc}(x,l,g) + \eta L_{conf}(x,c) + \beta L_{f-l}(x,c)) \quad (6)$$

其中，定位损失L_{loc}采用的是Smooth L1 Loss，分类损失L_{conf}采用的是多类别信息交叉熵，增加的惩罚项分类损失L_{f-l}采用Focalloss，N为与真实物体框所匹配的默认框数量，x为默认框与真实物体框的匹配结果，l为预测结果的位置信息，c为预测的类别标签，s为类别置信度；g为真实物体框个数，η和β参数用于调整两种分类损失的比例。

一种基于全局注意力的特征融合目标检测与识别方法

技术领域

[0001] 本发明涉及一种基于全局注意力的特征融合目标检测与识别方法，属于计算机视觉与智能信息处理领域。

背景技术

[0002] 目标检测是计算机视觉研究的主要领域之一，其主要任务是在一幅含有多目标物体的图像中，预测不同目标的类别标签与位置坐标。在实际生活中，目标检测已引起广泛关注，并已成功应用于许多领域，包括人机交互、人脸识别、计算摄影、自动驾驶、视频监控等各个方面。

[0003] 早期的检测方法利用手工制作的特征和分类器，特征表达能力非常有限。由于CNN的应用，基于CNN的检测器能有效地提取图像目标的特征，也使得AlexNet、GoogLeNet、ResNet和VGGNet等大型卷积神经网络模型得到了训练，实现了CNN强大的特征表达能力。基于深度学习的目标检测方法可分为两阶段检测方法和一阶段检测方法两类。两阶段检测方法将检测问题划分为两个过程，首先通过选择性搜索生成一组候选框，然后根据各候选区域的特征，采用卷积网络进行分类和回归，预测对象的位置和相应的类别标签。最具代表性的两阶段检测方法有R-CNN，FastR-CNN、FasterR-CNN等。以R-CNN方法为代表的两阶段检测方法虽然检测精度越来越高，但是其速度却遇到瓶颈，很难满足部分场景实时性的需求，因此出现了一种基于回归方法的一阶段检测方法。相较于两阶段检测方法，一阶段检测方法直接将目标框定位问题转化为回归问题，仅仅使用一个CNN网络预测不同目标的类别标签与位置坐标，在保证一定准确率的前提下，速度得到极大提升，经典的一阶段检测方法有YOLO和SSD等。虽然一阶段检测方法凭借高效率的优势近年来引起了更多关注，但是由于一阶段检测方法在小目标检测上有局限性，并且大多数现有方法为了提高精度引入复杂网络而牺牲了速度，为了克服性能和复杂性之间取舍的矛盾，本发明提出了一种基于全局注意力的特征融合目标检测与识别方法，在提升小目标物体检测效果的同时，平衡了检测的速度。

发明内容

[0004] 本发明提出了一种基于全局注意力的特征融合目标检测与识别方法，目的在于结合全局注意力模块和特征融合的方法得到表达能力更强的特征，在提高对小目标的检测效果的同时，平衡检测速度。

[0005] 本发明通过以下技术方案来实现上述目的：

[0006] (1)使用基准网络VGG-16作为特征抽取网络，末端辅以一系列卷积层，再结合空洞卷积RFB模块，初步得到多尺度的卷积特征层。

[0007] (2)采用多级特征融合方法，将初步得到的第1、2、3层特征进行融合到第1层特征图上，通过将深层和浅层的特征语义信息相结合，得到更加有效的特征。

[0008] (3)将融合得到的特征图和其他与特征层结合全局注意力模块，捕捉特征图像素

之间的关系，自适应地标定通道响应，从而提高特征图的表达能力。

[0009] (4)在多任务损失函数的基础上引入了Focal loss来调整正负样本的平衡。然后将上述处理后的特征进行分类和边框回归操作，通过训练，不断优化网络参数，再通过NMS过滤重复检测的边框得到最终的检测模型。

附图说明

[0010] 图1为本发明基于全局注意力的特征融合目标检测与识别方法网络框架图。

[0011] 图2为本发明基于多级特征融合模块结构图。

[0012] 图3为本发明基于注意力模块结构图。

具体实施方式

[0013] 下面结合附图对本发明作进一步说明：

[0014] 构建多级特征融合模块方法如下：

[0015] 本发明提出多级特征融合方法，将较深层特征融合到浅层特征上，使浅层特征得到更多的语义信息补充，最后得到感受野合适而又不缺乏语义信息的特征，从而更好的检测到小目标。构建多级特征融合模块如图2所示。

[0016] 首先，对第1层特征进行卷积操作，得到尺寸大小不变、通道为原来通道的1/3的特征，然后对第2、3层特征进行卷积和上采样操作，得到尺寸和第1层尺寸相同、通道为原第1层特征通道的1/3的特征，接着对经过处理的特征进行融合操作，得到感受野合适而又不缺乏语义信息的特征，用于更好的检测小目标。经过语义融合得到的特征图可以表示为

[0017] $$x_1 = ReLU\left\{W_k\left\{Concate\left[\mathcal{T}_i(x_1), \mathcal{T}_i(x_2)\right]\right\}\right\} \quad (1)$$

[0018] 其中x_i表示第i层特征图，$x_i \in R^{H\times W\times C}$，H、W和C分别表示特征图的长、宽和通道，$\mathcal{T}_i$表示对特征图进行卷积或上采样一系列操作，将$x_1$、$x_2$和$x_3$变换到指定尺寸和通道后进行Concate融合，接着经过卷积W_k和ReLU操作得到更有效的特征。

[0019] 构建全局注意力模块方法如下：

[0020] 全局注意力模块的结构图如图3所示。本发明中全局注意力模块分为三个部分，分别是Context Modeling、Transform和Fusion。

[0021] Context Modeling是上下文建模部分，采用1×1卷积w_k和Softmax等操作来获取注意力的权值，将全局上下文建模为所有位置特征的加权平均值，然后聚集全局上下文特征到每个位置的特征上，定义x为输入的特征图，特征图的宽和高分别为W和H，$\mathbf{x} = \{\mathbf{x}_i\}_{i=1}^{N_p}$，$x_i$和$x_j$分别表示某一位置的像素值，$x_i$经过Context Modeling得到的表达为

[0022] $$y_i = \sum_{j=1}^{N_p} \alpha_j x_j \quad (2)$$

[0023] 其中N_p为特征图的位置数量，$N_p = H \cdot W$，α_j用来计算位置i和所有可能关联的位置j之间的关系，获取全局上下文信息的权重，$\alpha_j = \frac{\exp(W_k x_j)}{\sum_{m=1}^{N_p}\exp(W_k x_m)}$，i表示该特征图中具体位置的

索引，j是所有可能的位置的索引，位置i和j两点之间的相似性关联函数通过$\exp(W_k x_j)$表征；W_k是1×1卷积操作，这里看作是一个线性转换矩阵。

[0024] Transform是特征转换部分，如图3所示，它通过卷积、Global pooling和ReLU等操作实现，用来捕获通道间的依赖关系，经过Transform模块得到的特征表示为

[0025] $s=x_c \cdot \sigma\{W_u \delta\{LN[W_r F_g(x_c)]\}\}$ (3)

[0026] 其中x_c是对ContextModeling模块的输出进行变换和卷积W_v操作得到的特征，δ是ReLU操作，σ是Sigmoid操作。接着在x_c的每个通道上执行全局平均池化，在空间维度上对特征进行压缩，使其具有全局的感受野，池化的过程F_g表示为

[0027] $$F_g(x_c)=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}x_c(i,j) \quad (4)$$

[0028] F_g操作将特征空间上所有点的信息平均成了一个值，表征在特征通道上响应的全局分布，因为要利用通道间的相关性，就需要基于通道的整体信息计算，增加了全局平均池化屏蔽空间分布的相关性而突出通道间的相关性。接着通过W_r卷积操作减少通道数从而降低计算量，加入LayerNorm作为正则化提高泛化性；然后通过W_u卷积操作恢复到原通道大小，以学习的方式为每个特征通道生成权重，来显式地建模特征通道间的相关性；最后经过Sigmoid激活函数层，获得0～1之间归一化的权重，通过乘法逐通道将归一化后的权重加权到原来的特征上，完成在通道维度上的对原始特征的重标定。综合来看，Transform以特征通道之间的关系为出发点，让网络利用全局信息，显式地建模特征通道之间的依赖关系，通过学习自动获取每个特征通道的重要性，选择性地增强重要的通道特征并抑制不重要的通道特征，从而达到特征通道自适应校准的目的。

[0029] Fusion是特征融合部分，用于将处理得到的全局上下文特征聚合到原始特征上，得到最终输出特征的表达式为

[0030] $z=x+s$ (5)

[0031] 优化的损失函数的模块如下：

[0032] 在多任务损失函数的基础上，加入了Focal loss惩罚项，对损失函数进行了改进，改善了单级探测器所面临的类别不平衡的问题，总目标损失函数由每个默认框的定位损失与分类损失的加权和构成，即

[0033] $$L(x,c,l,g)=\frac{1}{N}\left(L_{loc}(x,l,g)+\eta L_{conf}(x,c)+\beta L_{f\text{-}l}(x,c)\right) \quad (6)$$

[0034] 其中，定位损失L_{loc}采用的是Smooth L1 Loss，分类损失L_{conf}采用的是多类别信息交叉熵，分类损失$L_{f\text{-}l}$表示Focal loss，N为与真实物体框所匹配的默认框数量，x为默认框与真实物体框的匹配结果，l为预测结果的位置信息，s为类别置信度；g为真实物体框个数；η和β参数用于调整两种分类损失的比例，且η+β=1。L_{fl}表示Focal loss惩罚项，用于平衡正负样本。通过实验验证η=0.95，β=0.05时，可取较好的检测效果。L_{fl}计算公式为

[0035] $$L_{fl}=\begin{cases}-\alpha(1-\hat{y})^{\gamma}\log\hat{y}, y=1\\-(1-\alpha)\hat{y}^{\gamma}\log(1-\hat{y}), y=0\end{cases} \quad (7)$$

[0036] 其中，y是真实样本的标签(1为正样本，0为负样本)，$\hat{y}$是经过Sigmoid激活函数的

预测输出(数值在0-1之间)。平衡因子α用来平衡正负样本本身的数量比例不均,这里的两个参数α和γ协调来控制,本方法采用α=0.25,γ=2可以达到最好的实验效果。虽然只添加α可以平衡正负样本的重要性,但不能解决简单和难分样本的问题,因此针对难分样本的γ也必不可少,γ调节简单样本权重降低的速率,当γ为0时,即为交叉熵损失函数,当γ增加时,调整因子的影响也在增加。

[0037] 为了验证本发明所述基于全局注意力的特征融合目标检测与识别方法的有效性,在PASCAL VOC 2007和PASCAL VOC 2012两个数据集中开展实验。本文的实验的硬件环境为Inter(R)Xeon(R)CPUE5-2686的中央处理器,Nvidia GTX 1080Ti的显卡,16GB的RAM的PC机;软件环境为Ubuntu16.04.5系统,OpenCV和Pytorch深度学习开发框架,加速库为CUDA8.0和CUDNN6.0。采用VGG-16用作基础网络,采用SGD对得到的模型进行微调,学习率初始化为0.006,权重衰减为0.0005,动量为0.9,所有的卷积层使用“Xavier”方法进行初始化。检测精度的评价指标为mAP(mean Average Precision),检测时间性能的评价指标为FPS(Frames Per Second)。

[0038] 对于VOC 2007数据集,使用VOC 2007trainval和VOC 2012 trainval共16551张图像作训练集,用VOC 2007 test的4952张图像作测试集,所有这些图像都用类标签和真实边界框注释,通过迭代计算400个epoch,得到最终检测模型。实验结果由表1所示,当输入图像大小为300×300时,本发明的方法mAP为80.48%,比RFBNet300*高0.76%,领先于两阶段和YOLO,YOLOv2检测方法的同时,相较于SSD、RSSD和DSSD一阶段检测方法精度分别提高了6.16%、1.96%和1.86%;由于本发明只引入轻量级的计算,提升检测精度的同时降低了时间成本,速度达到81.7fps,比RFBNet300*略低。同样的,当输入图像尺寸为512×512时,与其他检测方法相比,精度和速度均有不同程度的提升,充分证明了本发明的有效性。

[0039] 对于VOC 2012数据集,使用VOC 2007 trainval和VOC 2012 trainval中的图像做训练集,用VOC 2012 test的10991张图像用于测试集,没有公共ground-truth边界框可用,所有方法的测试结果提交给PASCAL VOC的评估服务器评估。如表2所示,提供了每个类别的平均精度(AP)的详细比较,可以看到本发明增强了区分不同类别对象的模型能力,从而提高了大多数类别对象的检测准确性。

[0040] 表1 VOC 2007数据集实验结果

[0041]

Stage	Algorithm	Backbone	Input size	GPU	FPS	mAP（%）
Two-stage	Fast R-CNN[12]	VGG-16	~ 600 × 1000	Tian X	0.40	70.0
	Faster R-CNN[13]	ResNet-101	~ 600 × 1000	K40	7.00	73.2
One-stage	YOLO[14]	GoogleNet	448×448	Tian X	45.00	63.4
	YOLO v2[15]	DarkNet-19	544×544	Tian X	40.00	78.6
	SSD300[16]	VGG-16	300×300	Tian X	46.00	74.3
	SSD512[16]	VGG-16	512×512	Tian X	19.00	79.5
	RSSD300[17]	ResNet-101	300×300	Tian X	35.0	78.5
	RSSD512[17]	ResNet-101	512×512	Tian X	16.6	80.8
	DSSD321[18]	ResNet-101	321×321	Tian X	9.50	78.6
	DSSD513[18]	ResNet-101	513×513	Tian X	5.50	81.5
	RFBNet300*[19]	VGG-16	300×300	1080Ti	83.1	79.72
	RFBNet512*[19]	VGG-16	512×512	1080Ti	34.3	81.39
	Proposed300	VGG-16	300×300	1080Ti	81.7	80.48
	Proposed512	VGG-16	512×512	1080Ti	40.2	82.24

[0042] 表2 VOC 2012数据集实验结果

[0043]

Algorithm	mAP	AP（%）																			
	（%）	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster R-CNN[13]	73.8	86.5	81.6	77.2	58.0	51.0	78.6	76.6	93.2	48.6	80.4	59.0	92.1	85.3	84.8	80.7	48.1	77.3	66.5	84.7	65.6
YOLO[14]	57.9	77.0	67.2	57.7	38.3	22.7	68.3	55.9	81.4	36.2	60.8	48.5	77.2	72.3	71.3	63.5	28.9	52.2	54.8	73.9	50.8
YOLOv2[15]	73.4	86.3	82.0	74.8	59.2	51.8	79.8	76.5	90.6	52.1	78.2	58.5	89.3	82.5	83.4	81.3	49.1	77.2	62.4	83.8	68.7
SSD300[16]	75.8	88.1	82.9	74.4	61.9	47.6	82.7	78.8	91.5	58.1	80.0	64.1	89.4	85.7	85.5	82.6	50.2	79.8	73.6	86.6	72.1
DSSD321[18]	76.3	87.3	83.3	75.4	64.6	46.8	82.7	76.5	92.9	59.5	78.3	64.3	91.5	86.6	86.6	82.1	55.3	79.6	75.7	85.2	73.9
RSSD300[17]	76.4	88.0	83.8	74.8	60.8	48.9	83.9	78.5	91.0	59.5	81.4	66.1	89.0	86.3	86.0	83.0	51.3	80.9	73.7	86.9	73.8
RFB300*[19]	76.8	88.7	84.7	73.5	62.6	54.4	83.6	80.7	90.3	59.7	82.8	64.0	88.5	85.4	86.3	83.5	54.3	82.4	72.9	85.3	73.2
Proposed300	77.3	88.4	85.3	74.6	62.9	55.5	83.9	80.5	91.1	61.8	83.0	63.9	89.0	85.9	86.4	84.0	55.2	82.1	72.7	85.9	74.1

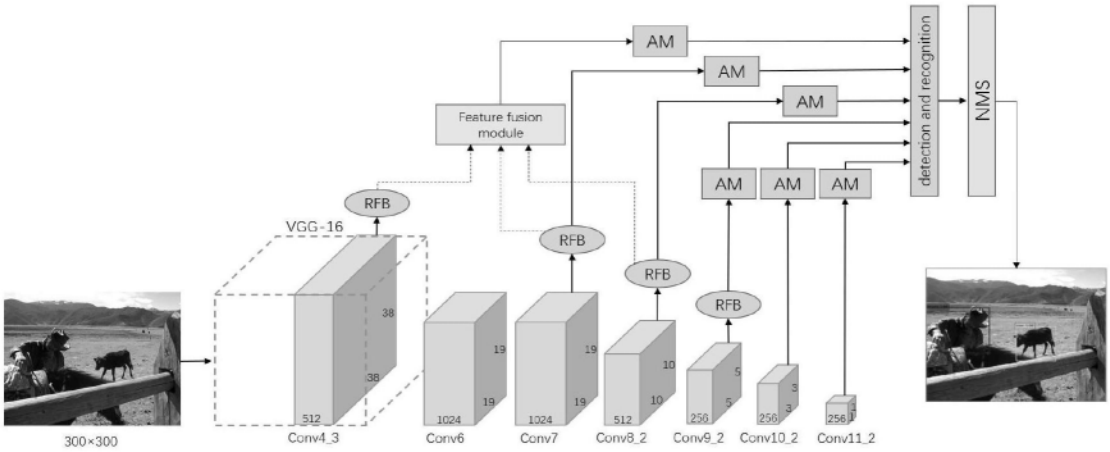

AM
AM
AM
AM
AM
AM
detection and recognition
NMS
Feature fusion module
RFB
RFB
RFB
RFB
VGG-16
300×300
Conv4_3
Conv6
Conv7
Conv8_2
Conv9_2
Conv10_2
Conv11_2

图1

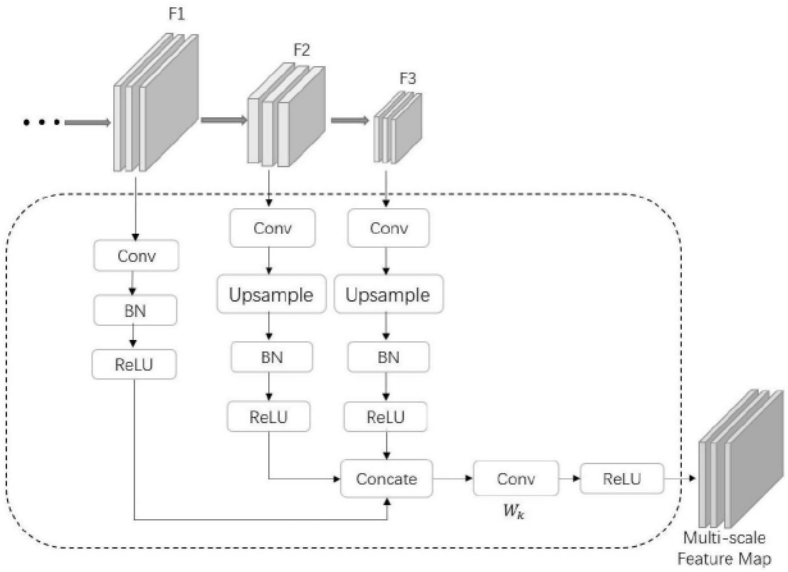

F1
F2
F3
Conv
BN
ReLU
Conv
Upsample
BN
ReLU
Conv
Upsample
BN
ReLU
Concate
Conv
ReLU
W_k
Multi-scale
Feature Map

图2

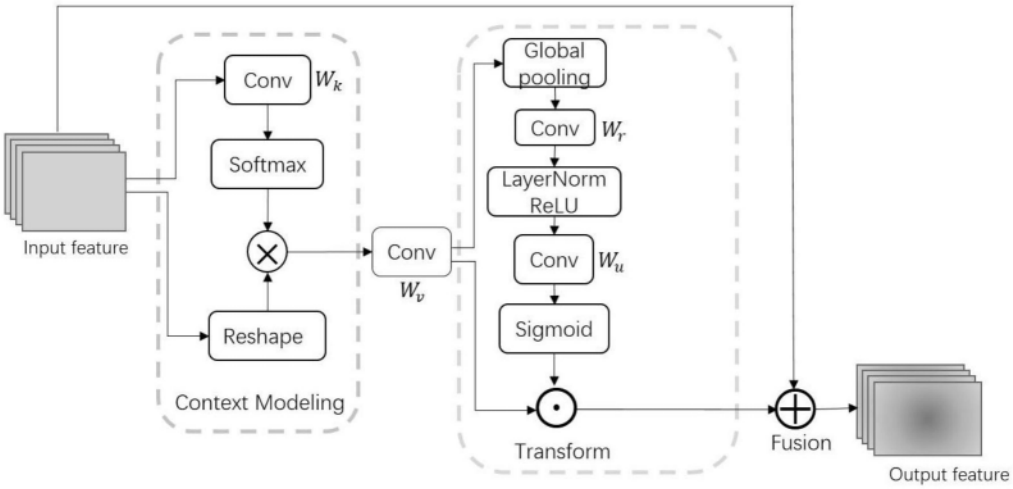

Input feature
Conv
W_k
Softmax
Reshape
Context Modeling
Conv
W_v
Global pooling
Conv
W_r
LayerNorm ReLU
Conv
W_u
Sigmoid
Transform
Fusion
Output feature

图3