计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):107−112 [doi: 10.15888/j.cnki.csa.006755]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

面向车辆检测的扩张全卷积神经网络①

程雅慧[1,2,3], 蔡　烜[4], 冯　瑞[1,2,3]

[1](复旦大学 计算机科学技术学院, 上海 201203)
[2](上海视频技术与系统工程研究中心, 上海 201203)
[3](复旦大学 上海市智能信息处理实验室, 上海 201203)
[4](物联网技术研发中心, 上海 201204)

摘　要: 近年来, 深度学习方法被广泛用来解决车辆检测问题并取得了显著的成果, 然而, 当车辆尺寸较小时, 当前深度学习算法的检测丢失率仍然很高. 为了解决这个问题, 本文提出了一种基于组合目标框提取结构的扩张全卷积神经网络 (Dilated Fully Convolutional Network with Grouped Proposals, DFCN-GP). 具体提出了一种结合低层特征和高层特征的组合网络模型用于生成目标框, 其中低层特征对小目标更加敏感. 此外, 为保留更多的细节信息, 基于扩张卷积思想, 增加了网络最后一层卷积层的大小和感受野, 用于目标框的提取和车辆检测. 通过控制变量的对比试验, 对基于组合方式的目标框提取网络和扩张卷积层的有效性进行了验证. 本文提出的算法模型在公开数据集 UA-DETRAC 上性能优异.

关键词: 机器视觉; 车辆检测; 组合网络模型; 扩张全卷积神经网络

引用格式: 程雅慧,蔡烜,冯瑞.面向车辆检测的扩张全卷积神经网络.计算机系统应用,2019,28(1):107–112. http://www.c-s-a.org.cn/1003-3254/6755.html

Dilated Fully Convolutional Network with Grouped Proposals for Vehicle Detection

CHENG Ya-Hui[1,2,3], CAI Xuan[4], FENG Rui[1,2,3]

[1](School of Computer Science, Fudan University, Shanghai 201203, China)
[2](Shanghai Engineering Research Center for Video Technology and System, Shanghai 201203, China)
[3](Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, Shanghai 201203, China)
[4](Internet of Things Technology Research and Development Center, Shanghai 201204, China)

Abstract: Although deep learning based vehicle detection approaches have achieved remarkable success recently, they are still likely to miss comparatively small-sized vehicle. To address this problem, we propose a novel Dilated Fully Convolutional Network with Grouped Proposals (DFCN-GP) for vehicle detection. Specifically, we invented a grouped network structure to combine feature maps from both lower and higher level convolutional layers for the generation of object proposal and focusing more on lower level features, which are more sensitive to discovering small object. In addition, we increase the size and reception field of the feature map in the last convolutional layers to keep more detailed information via dilated convolution, which is used in both object proposal and vehicle detection sub-networks. In the experiment, we conducted ablation studies to demonstrate the effectiveness of the grouped proposals and dilated convolutional layer. We also show that the proposed approach outperforms other state-of-the-art methods on the UA-DETRAC vehicle detection.

Key words: machine vision; vehicle detection; grouped region proposals; dilated convolutional networks

① 收稿时间: 2018-07-31; 修改时间: 2018-08-27; 采用时间: 2018-09-05; csa 在线出版时间: 2018-12-26

车辆检测[1]在计算机视觉领域是一个很重要的研究课题, 如智能交通, 自动驾驶等. 近年来, 业内已提出了很多基于深度学习的车辆检测算法, 实际应用表明, 基于深度学习的方法比传统手工提取特征方法更加有效.

基于区域的卷积神经网络 (Region-based Convolutional Neural Network, RCNN) 是一种典型的基于深度学习的目标检测算法, 在车辆检测领域被广泛使用. 基于 RCNN 框架, 通过卷积神经网络提取给定的候选目标区域特征, 用于预测目标类别和目标位置. Faster-RCNN 算法通过在 RCNN 最后一层卷积层提取出的特征同时用于候选目标框的生成和最终的目标检测.

研究发现, 随着网络层数的不断增加, 卷积神经网络 (Convolution Neural Network, CNN) 能很好地学习低层特征和高层语义特征. 高层卷积层在分类任务方面比定位任务具有更好的性能, 有学者[2,3]试图结合多种低/高层卷积层用于目标检测, 尽管这些方法不同程度上改善了目标检测性能, 但是在有不同尺度的车辆检测方面, 多个不同尺寸的车辆检测精度无法达到平衡. 特别地, 相对于大尺寸目标, 这些方法在小尺寸目标检测方面表现不尽人意.

本文提出了一种适用于小尺寸车辆的检测算法, 即基于组合目标框提取结构的扩张全卷积神经网络 (Dilated Fully Convolutional Network With Grouped Proposals, DFCN-GP). 图 1 为 DFCN-GP 的算法框架, 算法包括用于图像特征提取的扩展卷积网络 (Dilated Convolutional Network, DCN)、组合目标框提取网络 (Grouped Region Proposal Network, GRPN), 以及用于车辆检测的全卷积网络 (Fully Convolutional Network, FCN) 三部分.

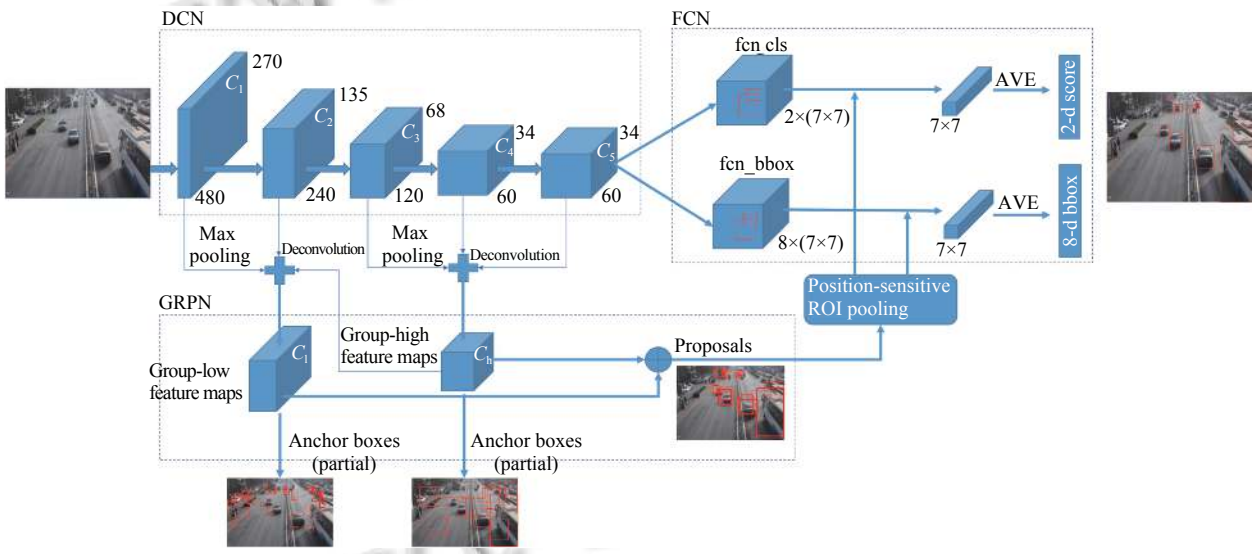

图 1 DFCN-GP 算法框架图

本文方法的主要贡献为: 1) 提出了两组组合方式, 即混合从低层到高层卷积层的特征图, 并将两组混合特征应用到区域目标框提取网络, 同时生成候选车辆区域. 与现有方法不同, 这种组合结构更多地聚焦低层卷积层的特征. 2) 将低层和高层特征产生的目标候选框进行融合, 有利于收集不同尺度的候选目标框, 降低目标丢失率. 3) 将最后一层卷积层改成扩张卷积层并把步长改为 1, 使最后一层卷积层的特征图拥有更大的尺寸和感受野, 可同时用于目标框提取网络和车辆检测网络, 进而在候选目标框的生成和最终目标检测中更多的细节信息. 上述改进可使模型能有效检测图像中较小的车辆目标.

在实验部分, 将本文提出的网络模型与现有最新的网络模型在 UA-DETRAC 车辆数据集上进行对比. 实验结果显示, 本文方法取得了 71.56% mAP, 超越了现有最好方法的结果. 此外本文提出的网络模型同样通过控制变量对比试验, 证明了组合特征提取方式和扩张卷积设置的有效性.

本文论文剩余部分安排如下: 第 1 节讨论了提出的算法模型, 第 2 节对本文算法进行了实验并对结果

进行了分析, 第 3 节为总结全文.

1 算法原理

本文提出一种用于车辆检测的端到端学习网络(DFCN-GP), 可从不同的卷积层级中实现候选目标框的提取和车辆检测. 如图 1 所示, 网络的前向路径包括以下几个步骤:

Step1. 图像特征提取. 利用 Resnet-101 和扩张卷积层提取输入图像的特征, 用于后续目标框的提取和车辆检测.

Step2. 区域候选目标框提取. 提出一种两组组合方式, 即从不同卷积层生成候选目标框.

Step3. 车辆分类和定位. 利用多任务学习方式, 在最后一层扩张卷积层同时判别候选框是否为车辆以及估计车辆的位置.

1.1 图像特征提取

通过 DCN 从输入图像中提取车辆信息特征. 具体地, 本文的特征提取网络基于改进的 ResNet-101[4]结构. 与传统卷积神经网络训练不同, ResNet 利用残差结构实现更为有效的模型训练, 在图像分类和目标检测任务中性能优异.

如图 1 所示, 本文的特征提取网络包括 5 个残差结构, 分别标注为$\{C_1,C_2,C_3,C_4,C_5\}$. 根据原生 ResNet-101 的设置, 每个残差块会连接步长为 2 的卷积层. 因此, 原生 ResNet-101 网络中每个残差块的特征图步长分别是$\{2,4,8,16,32\}$, 即给定$X\times Y$大小的输入图像, 则最后的卷积层的特征图C_5, 输出大小为$X/32\times Y/32$. 当C_5用于目标框提取和车辆检测时, C_5的尺寸相对于原始图像的细节信息过小, 而且容易丢失小目标.

为了解决上述问题, 考虑从一下两个方面对网络进行改进: 1) 改变C_4后的卷积层步长, 将步长 2 改为步长 1, 进而增加最后一层卷积层的特征图大小. 通过这样方式, C_5直接接在C_4后面, 不加任何下采样处理. 2) 基于文献[5]的思路, 应用扩张卷积的思想, 增加C_5的感受野, 最终网络的 5 个残差块的步长变为 $\{2,4,8,16,16\}$.

1.2 区域候选目标框提取

在此步骤中, 算法从提取的特征图中产生候选的目标框, 这个过程由区域目标框提取 (Region Proposal Network, RPN) 这样的子网络完成. 正如文献[6]中讨论的那样, 较高层的特征图包含更多的高级语义信息, 而较低层的特征图包含更多的细节信息. 为了更好地发现图像中不同尺度的车辆目标, 本文旨在探索不同级别的特征图所包含的信息以及更多地聚焦在小目标检测上. 为了这个目的, 本文发明一个基于组合的区域目标框提取网络, 用来结合两组卷积特征图. 这两个组分别被命名为 group-high 和 group-low. 在每个组中, RPN 网络在两个不同尺度级别上同时生成固定框 (anchors).

在 group-high 中, 将三个不同的特征图C_3, C_4和C_5链接起来. 具体通过添加一个最大池化操作 (步长为 2) 链接在C_3的后面, 添加一个1×1的卷积层分别接在C_4和C_5的后面. 接着将这些输出的特征图链接到一起 (相同的像素大小), 形成融合产生的特征图, 记为C_h. C_h的大小为34×60. 应用 RPN 于C_h产生若干候选目标框. 在 group-high 中, anchors 的尺度设置为$\{32^2,64^2,128^2,256^2,512^2\}$, 长宽比设置为$\{1:2,2:1,1:1\}$.

在 group-low 中, 结合C_1, C_2和 group-high 中产生的融合特征C_h, 将融合后的特征标记为C_l. 具体通过应用最大池化操作 (步长为 2) 于C_1, 上采样操作 (步长为 4) 于C_h, 以及1×1的卷积层于C_2, 将他们链接在一起 (相同的像素大小135×240). 将另一个 RPN 网络应用于C_1上来产生候选目标框. 在这个组中, anchors 的尺寸设置为$\{16^2,32^2\}$. 长宽比设置为$\{1:2,2:1,1:1\}$.

上述用于目标框提取的网络模型对现有方法进行了优化和提升: 原生 RPN 网络仅从最后一层卷积层C_5上计算目标框, 同时也使用了多尺度的 anchors. 有研究[2,3,7]提出结合C_1, C_3和C_5形成特征图, 即超-特征 (Hyper Feature), 然而, 尽管 Hyper Feature 结合了低层和高层的信息, 但对车辆检测, 尤其是对小目标车辆的检测性能不佳, 本文提出的“两组”策略在车辆检测任务中会带来更高的召回率.

1.3 车辆分类和定位

在前述步骤的基础上, 可通过 FCN 判别提取框是否为车辆, 并估计估计车辆的准确位置坐标. 为此, 模型首先通过 PS-ROI pooling[8]操作在C_5上裁剪出感兴趣区域, 然后再接上两个分支的全卷积网络. 第一个分支实现分类, 即对裁剪出的区域打分. 第二个分支实现位置预测, 即对有目标的区域生成一个 8 维的向量 (前景/背景).

在模型训练阶段, 本文的目标是最小化多任务损失函数, 包括分类和定位两个部分. 具体为给定标签值和预测值$\{s,s^*\}$和候选框的补偿值$\{t,t^*\}$, 目标检测损失函数定义如下:

$$L(s,s^*,t,t^*)=L_{cls}(s,s^*)+\lambda L_{reg}(t,t^*) \quad (1)$$

在公式 (1) 中$L_{cls}(s,s^*)$是分类任务的交叉熵损失, $L_{reg}(t,t^*)$是候选框回归任务中的平滑L_1损失. t表示 4 维向量(t_x,t_y,t_w,t_h), 代表位置坐标中标签值与 anchor 之间的差值. t值计算如下:

$$\begin{aligned} t_x&=(G_x-A_x)/A_w\\ t_y&=(G_y-A_y)/A_h\\ t_w&=\log(G_w/A_w)\\ t_h&=\log(G_h/A_h) \end{aligned} \quad (2)$$

在公式 (2) 中, $A=(A_x,A_y,A_w,A_h)$由 anchor 计算得到, 而$G=(G_x,G_y,G_w,G_h)$是代表标签的目标框. 根据文献[9]中的设置, 平衡λ的缺省值为 1.

相对于常规 Resnet-101 模型,本文模型直接预测车辆是通过改良的具有大尺寸和大感受野的C_5特征图. 我们发现在改良过的C_5上预测会比 Hyper Feature 效果更好.

2 实验部分

2.1 数据集和实验设置

在实验部分, 首先设计控制变量对比实验验证本文网络各组成部分的有效性, 然后将本文模型与其他典型模型进行对比. 算法模型在 UA-DETRAC[10]数据集上训练模型, 该数据集由 10 个视频组成, 视频由 Cannon EOS 550D 的摄像机在中国北京和天津两座城市共 24 个不同的位置录制而成. 视频每秒 25 帧且像素为 960×540, 数据集总共超过 14 万帧. 数据集中共有 8250 个车辆, 121 万个车辆矩形框被标注.

对输入数据, 首先将训练图片归一化, 将图片短边归一化为 540. 提取图像特征的共享卷积层初始参数设置为在 Imagenet 上训练的 Resnet-101 参数, 其余层的初始参数采用 0 均值, 0.01 方差的高斯随机数.

在区域目标框提取网络训练阶段, 本文将与任何标签框面积重叠比率 (Intersection-Over-Union, IOU) 大于 0.7 的候选提取框设置为含有目标的候选框. 将与任何标签框的 IOU 小于 0.3 的候选提取框设置为不含目标的框.

在目标检测网络训练部分, 本文将与任何标签框的 IOU 大于 0.6 的候选提取框设置为含目标的候选框. 将与任何标签框的 IOU 小于 0.6 的候选提取框设置为不含目标的框. 在测试阶段, 目标检测网络部分应用 NMS 来合并有重叠的候选框.

在训练过程中, 将批大小设置为 128, 总迭代次数设置为 9 万轮. 学习率初始值为 1e–3, 每 3 万轮迭代乘以系数 0.1. 在算法效率方面, 本文的模型在测试阶段处理一张图像 (960×540) 需 0.4 s.

2.2 控制变量对比试验

对于区域提取目标框网络和检测网络来说, 输入设置对最终检测效果至关重要. 下面针对不同特征图的结合方式设置了几组控制变量对比试验, 并讨论其效果.

2.2.1 区域提取目标框网络输入设置

本文尝试了不同卷积层的结合对目标框提取的影响. 表 1 为使用不同卷积层作为 RPN 的输入对应的提取框召回率, 以及其最终的检测率. 目标包括小/中/大车辆尺寸. 最后一列为最终的检测精度 (Average Precision, AP). 行 (a) 是将C_1,C_3,C_5结合在一起作为 RPN 的输入. 通过设置 6 种提取框的尺度, 分别为 $\{16^2,32^2,64^2,128^2,256^2,512^2\}$. 从结果可以看出, 小目标的召回率明显小于中等目标和大目标的召回率. 行 (b) 是将C_3,C_4,C_5结合在一起作为 RPN 的输入, 同样也是设置 6 种提取框的尺度. 从结果可以发现大目标的召回率明显提升, 但小目标及中等目标的召回率大幅度下降. 和行 (b) 相比, 行 (c) 简单地结合低层特征C_1,C_2来预测小目标和中等目标框. 本文对于高层的融合特征图设置 5 种目标框尺度为$\{32^2,64^2,128^2,256^2,512^2\}$, 对于低层融合特征设置 2 种目标框尺度为$\{16^2,32^2\}$. 从结果可以看出, 中等和小目标的召回率有了轻微的提升, 但是没有融合高层语义特征使得小目标并不能正确的分类出来. 最终在行 (d) 中, 本文提供了自己提出网络的结果. 同样对高层融合特征C_h设置 5 种目标框尺度, 对低层融合特征设置 2 种目标框尺度. 最终发现本文的融合策略可以大幅度改善小目标和中等目标的召回率, 并且维持大目标的召回率基本不变. 本文提出的融合方法在目标框提取中对于寻找各种尺寸的目标框达到一个很好的平衡. 另外, 最终本文方法的检测精度 (行 (d)) 也同样高于其它模型.

表 1 不同卷积层特征图的结合方法作为 RPN 网络的输入在 UA-DETRAC 验证数据集上目标的召回率

输入卷积层	R^{100}_{small}	R^{100}_{medium}	R^{100}_{large}	AP
(a) $C_1-C_3-C_5$	69.3	81.4	84.4	85.3
(b) $C_3-C_4-C_5$	57.3	78.9	**90.0**	84.2
(c) $C_1-C_2+C_3-C_4-C_5$	60.8	79.1	89.0	84.5
(d) $C_1-C_2-C_h+C_3-C_4-C_5$	**70.3**	**83.9**	88.6	**87.4**

2.2.2 全卷积检测网络的输入设置

这部分实验了不同卷积层的特征会怎样影响最终

的检测结果. 平均检测精度 (AP) 结果如表 2 所示. 行 (a–d) 直接在单个卷积层C_1,C_3或C_5上检测目标. 从表中可以发现, 利用高层特征C_5(行 (c, d)) 作为输入比利用低层特征特征作为输入要取得更好的检测效果. 这样说明相比于低层特征, 高层特征在检测任务中更加有效. 注意到行 (d) 是本文提出网络的设计, 和行 (c) 不同点仅在于C_5的特征图尺寸和感受野要大. 行 (d) 比行 (c) 结果更好证明了将C_5步长改为 1 以及扩张卷积思想的引入是非常有效的. 行 (e) 也例举了C_1−C_3−C_5 (Hyper Feature) 融合特征作为检测网络输入的结果. 结果显示改良的C_5(×16) 能更好地检测目标, 同时也说明在检测网络中引入低层特征可能会带来更多噪声, 并不适合检测任务.

表 2 经过 ROI 池化操作后不同的卷积特征图作为最终全卷积检测网络的输入在 UA-DETRAC 验证数据集上平均精度

输入卷积层	提取框	*AP* (%)
(a) C_1	*GRPN*(×16)	59.8
(b) C_3	*GRPN*(×16)	78.3
(c) C_5(×32)	*GRPN*(×32)	84.1
(d) C_5(×16)	*GRPN*(×16)	87.4
(e) C_1+C_3+C_5(×16)	*GRPN*(×16)	83.2

2.3 和现有模型比较

将本文的模型与其他现有目标检测模型对比 (UA-DETRAC 测试集). 用来对比的检测模型包括 DPM[11], ACF[12], R-CNN[13], Faster-RCNN[14], CompACT[15]和 EB. ACF 和 DPM 是经典的目标检测方法. ACF 基于一种加速框架而 DPM 基于 SVM(支持向量机). R-CNN 和 Faster-RCNN 都是基于区域提取的目标检测卷积神经网络. CompACT 通过一种级联的方式检测目标. EB 是基于 Faster-RCNN 的改良算法, 增加一个额外的精调网络.

表 3 列出了不同方法在不同子测试集上的检测精度. 之前最好的检测网络是 EB(获 ICME2017 最佳论文), 通过对比发现, 本文方法在全集上的 AP 为 71.56%. 注意到本文提出的方法在不同子类别上的检测精度也是最高的. 图 2 展示了不同检测方法的精度-召回率曲线图. 这同样说明相比于其他方法, 本文方法具有明显的优越性. 图 3 图形化展示了一些方法在 UA-DETRAC 数据集上的检测结果, 可以看出本文方法对于不同尺度的车辆以及不同的天气条件下都具有出色的检测效果.

表 3 和其他先进的方法在 UA-DETRAC 测试集上的检测结果

方法	全集	容易	中等	难	多云	夜晚	雨天	晴天
DPM	25.70	34.42	30.29	17.62	30.91	31.77	25.55	24.78
ACF	46.35	54.27	51.52	38.07	35.29	66.58	37.09	58.30
R-CNN	48.95	59.31	54.06	39.47	39.32	67.52	39.06	59.73
Faster-RCNN	58.45	82.75	63.05	44.25	66.29	69.85	45.16	62.34
CompACT	53.23	64.84	58.70	43.16	46.37	71.16	44.21	63.23
EB	67.96	89.65	73.12	54.64	72.42	73.93	53.40	83.73
DFCN-GP (ours)	**71.56**	**93.51**	**78.00**	**55.62**	**74.84**	**79.37**	**57.22**	**84.92**

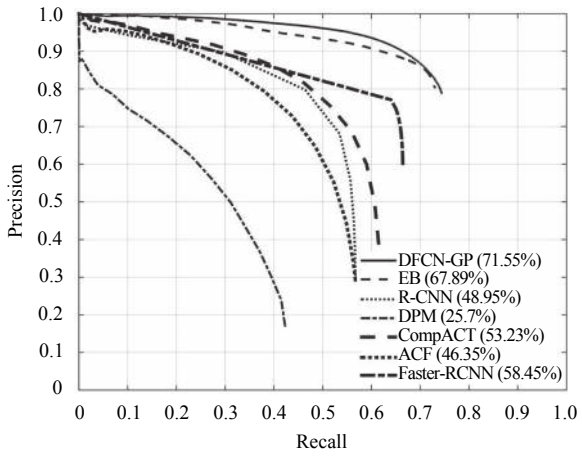

图 2 对比检测算法在 UA-DETRAC 测试集上的精度-召回率曲线

图 3 本文提出的车辆检测算法在 UA-DETRAC 测试集上的结果 (红色框标出)

3 总结

本文提出了一种新的车辆检测算法 DFCN-GP. 通过在目标框提取和车辆检测中引入了两组卷积层特征图和扩张卷积层, 实现了小目标车辆检测精度的大幅度提升. 控制变量对比试验和与先进模型对比的实验结果验证了本文算法的有效性.

参考文献

1 董春利, 董育宁. 基于视频的车辆检测与跟踪算法综述. 南京邮电大学学报 (自然科学版), 2009, 29(2): 88–94. [doi: 10.3969/j.issn.1673-5439.2009.02.018]

2 Kong T, Yao AB, Chen YR, *et al*. HyperNet: Towards accurate region proposal generation and joint object detection. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 845–853.

3 Wang L, Lu Y, Wang H, *et al*. Evolving boxes for fast vehicle detection. Proceedings of 2017 IEEE International Conference on Multimedia and Expo. Hong Kong, China. 2017. 1135–1140.

4 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA. 2016. 770–778.

5 Yu F, Koltun V, Funkhouser T. Dilated residual networks. arXiv preprint arXiv: 1705.09914, 2017.

6 Ghodrati A, Diba A, Pedersoli M, *et al*. DeepProposal: Hunting objects by cascading deep convolutional layers. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 2578–2586.

7 Ranjan R, Patel VM, Chellappa R. HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv preprint arXiv: 1603.01249, 2016.

8 Dai JF, Li Y, He KM, *et al*. R-FCN: Object detection via region-based fully convolutional networks. Advances in Neural Information Processing Systems 29. Springer. 2016. 379–387.

9 Girshick R. Fast R-CNN. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 1440–1448.

10 Wen LY, Du DW, Cai ZW, *et al*. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. arXiv preprint arXiv:1511.04136, 2015.

11 Felzenszwalb PF, Girshick R, McAllester D, *et al*. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627–1645. [doi: 10.1109/TPAMI.2009.167]

12 Dollár P, Appel R, Belongie S, *et al*. Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532–1545. [doi: 10.1109/TPAMI.2014.2300479]

13 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, OH, USA. 2014. 580–587.

14 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada. 2015. 91–99.

15 Cai ZW, Saberian M, Vasconcelos N. Learning complexity-aware cascades for deep pedestrian detection. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 3361–3369.