DOI:10.11918/201904144

结合双重注意力机制的遮挡感知行人检测

周大可[1,2], 宋 荣[1], 杨 欣[1]

(1.南京航空航天大学 自动化学院, 南京 211100; 2.江苏省物联网与控制技术重点实验室(南京航空航天大学), 南京 211100)

摘 要: 针对行人检测算法在交通场景下应用时的遮挡问题,提出一种结合双重注意力机制的遮挡感知行人检测算法。以 RetinaNet 作为基础框架,在回归和分类支路分别添加空间注意力和通道注意力子网络,增强网络对于行人可见区域的关注;同时引入行人可见边界框信息对传统的回归损失函数进行优化,使其能够随着遮挡程度自适应地调节预测框贡献的权重。在 Caltech 和 CityPerson 数据集上的实验结果表明:相较于 RetinaNet 等 8 种先进算法,该方法具有较好的鲁棒性和检测精度,尤其是严重遮挡情况下,该算法的对数平均漏检率仅为 45.69%,小于其他算法 12%以上;此外,该算法能够实现准实时检测,在 Caltech 和 CityPerson 上的检测速度分别为 11.8 帧/s 和 10.0 帧/s。所提出的双重注意力机制和遮挡感知回归损失函数的检测方法具有可行性和有效性,对于遮挡行人的处理有显著优势。

关键词: 行人检测;卷积神经网络;注意力机制;遮挡;实时

中图分类号: TP399 **文献标志码**: A **文章编号**: 0367-6234(2021)09-0156-08

Occlusion-aware pedestrian detection combined with dual attention mechanism

ZHOU Dake[1,2], SONG Rong[1], YANG Xin[1]

(1.College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 211100, China;
2.Jiangsu Key Laboratory of Internet of Things and Control Technologies (Nanjing University of Aeronautics and Astronautics), Nanjing 211100, China)

Abstract: To address the occlusion problem of pedestrian detection algorithm when applied in traffic scenarios, this paper presents an occlusion-aware algorithm combined with dual attention mechanism for pedestrian detection. Based on the RetinaNet framework, the spatial-wise attention mechanism and channel-wise attention mechanism were utilized in regression and classification branches respectively, guiding the detector to pay more attention to the visible parts of pedestrians. Moreover, visible bounding box information of pedestrians was introduced to optimize the traditional regression loss function, so that it can adaptively adjust the weights of predicted boxes according to the degree of occlusion. Experiments on Caltech and CityPerson datasets show that the proposed algorithm had better robustness and higher accuracy than other eight advanced algorithms such as RetinaNet. Especially in the case of heavy occlusion, the log-average miss rate of the proposed algorithm was only 45.69%, which was 12% lower than those of other algorithms. Furthermore, the proposed algorithm could detect pedestrians in quasi-real-time. It processed 11.8 frames per second on Caltech dataset and 10.0 frames per second on CityPerson dataset. The detection methods of dual attention mechanism and occlusion-aware regression loss function proposed in this paper are feasible and effective, and have significant advantages for the processing of occluded pedestrians.

Keywords: pedestrian detection; convolutional neural networks; attention mechanism; occlusion; real-time

行人检测作为目标检测领域的一个重要研究方向,一直受到研究者们的普遍关注,目前已经对智能交通、智能辅助驾驶和视频监控等领域产生了深入的影响[1]。传统的行人检测方法,如 HOG(histogram of oriented gradient)[2]、DPM(deformable parts model)[3]和 ACF(aggregate channel feature)[4]等,都是通过手工设计或特征聚合来获得行人特征。随着 2012 年 AlexNet[5]在图像分类任务中的重大突破,利用卷积神经网络 CNN(convolutional neural networks)自主学习特征提取过程从而代替传统手工设计是目前的主要研究方向[6]。根据检测机制的不同,基于卷积神经网络的目标检测方法主要分为两类:一是两阶段方法,以 Faster R-CNN[7]为例,其主要思路是采用级联的方式,在生成候选目标区域的基础上进一步判断边界框的类别和位置。另一类则是单阶段方法,以 YOLO(you only look once)[8]和 SSD(single shot multibox detector)[9]为例,其思路是用一个卷积神经

收稿日期: 2019-04-17
基金项目: 国家自然科学基金(61573182);南京航空航天大学研究生创新基地(实验室)开放基金(kfjj20180319)
作者简介: 周大可(1974—),男,副教授,硕士生导师
通信作者: 周大可,dkzhou@nuaa.edu.cn

网络直接回归出边界框的位置和类别。

卷积神经网络的引入提升了行人检测算法性能，但遮挡问题仍然是行人检测中的一个主要难点[10-13]。文献[10]通过一种联合学习方式建模不同的行人遮挡模式，但其检测框架复杂且无法穷尽所有的情况；文献[11]设计新的损失函数，使预测框在不断逼近目标真实框的同时远离其他的真实框，这种方法对遮挡的处理更为灵活，实现也更加简单；文献[12]将前述的两种思路相结合，提出部件遮挡感知单元和聚集损失函数来处理行人遮挡问题；文献[13]通过引入新的监督信息（行人可见区域边界框）来处理遮挡，其思路是用两个分支网络分别回归行人的全身框和可见区域的边界框，最终融合两个分支的结果来提升检测性能。

注意力机制源于对人类视觉的研究，在计算机视觉的各种任务（如图像分类、检测和分割等）中均有广泛的应用[14]。常见的注意力机制有两种类型：一是空间注意力机制[15]，即通过网络学习来自适应地调节特征图中每个元素的权重；二是通道注意力机制[16]，即利用网络来调节特征图中不同通道的权重。利用注意力机制可以加强网络对行人可见区域特征的关注，进而改善算法的遮挡处理能力。文献[17]利用预训练的行人姿态估计模型生成的部件热图作为监督信息指导通道注意力机制的学习，有效提高了遮挡行人的检测效果，但其仅使用了单一的通道注意力机制且需要额外的网络来生成监督信息，检测框架复杂。

本文以基于回归的检测方法 RetinaNet[18] 为基础，针对行人检测的两个子任务（分类和定位），在不同的支路分别采用空间和通道注意力机制，同时引入行人边界框作为监督信息，简单有效地指导两种注意力机制的学习。此外，利用行人可见区域边界框设计新型的可感知遮挡的回归损失函数，进一步提高了算法对遮挡的鲁棒性。

1　结合注意力机制的遮挡感知行人检测

1.1　网络整体结构

本文方法的基本框架采用 RetinaNet，主要由 3 个部分组成，分别是 Resnet[19] 主干网络、FPN[20]（feature pyramid network）特征金字塔融合模块、以及结合双重注意力机制的卷积预测模块，网络整体结构如图 1 所示。

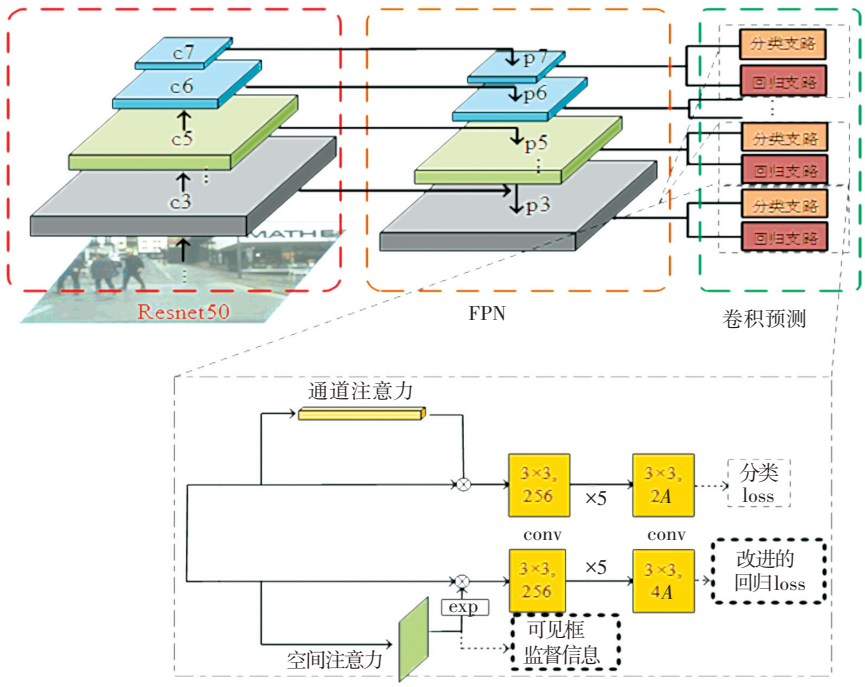


图 1　本文算法整体网络结构示意

Fig.1　Overall network structure of proposed algorithm

Resnet 是目前主流的特征提取主干网络之一，其通过“捷径”将前后层直接相连，从而使网络更加容易拟合恒等映射。Resnet 可以改善网络深度增加带来的模型训练困难、性能提升较小的问题，即“退化”现象。本文提取特征的主干网络采用 Resnet50，其具体结构参数见表 1。

FPN 是一种 U 型网络结构，其通过融合生成的特征金字塔，有效结合深浅层不同维度的特征表达，并且在不同层独立预测不同尺度的行人。如图 1 所示，自上至下的卷积层 c5、c4、c3 分别在采样之后与

下层逐层融合,得到 p5、p4、p3。p6 和 p7 即 c6 和 c7,在 c5 的基础上分别通过一次和两次 3×3 卷积得到。多层预测可以更好地处理行人远近导致的尺度问题。

表 1 Resnet50 结构

Tab.1 Structure of Resnet50

层名	卷积核参数	特征图分辨率/像素
conv1	7×7, 64	1 200×900
max pool	3×3	600×450
conv2_x	$\begin{pmatrix}1\times1, 64\\3\times3, 64\\1\times1, 256\end{pmatrix}\times3$	300×125
conv3_x	$\begin{pmatrix}1\times1, 128\\3\times3, 128\\1\times1, 512\end{pmatrix}\times4$	150×62
conv4_x	$\begin{pmatrix}1\times1, 256\\3\times3, 256\\1\times1, 1\,024\end{pmatrix}\times6$	75×31
conv5_x	$\begin{pmatrix}1\times1, 512\\3\times3, 512\\1\times1, 2\,048\end{pmatrix}\times3$	37×15

卷积预测模块包含分类支路和回归支路,分类支路主要负责区分前景与背景,其通过多个卷积核大小为 3×3,输出通道数为 256 的卷积层对 p3~p7 进行卷积,最终以通道数为 $K \times A$ 的 3×3 卷积输出类别概率。其中 K 为类别数目,本文中设为 2,即仅前景和背景两个类别,A 表示输出特征图中每个网格的先验边界框数目,本文中为 9。回归支路除了尾部输出卷积层以外结构均与分类支路相同,在此不再赘述。尾部输出卷积层需输出预测框相对于预设框的偏移程度,通过通道数为 $4A$ 的 3×3 卷积实现,4 表示框的偏移量 dx、dy、dw、dh。

本文在 RetinaNet 的基础上对卷积预测模块的分类支路和回归支路分别增加注意力机制子网络,同时引进行人可见框信息对传统的回归损失函数进行优化,如图 1 所示。除了以上两点改进之外,本文网络所有参数设定均保持与基准方法相同。

1.2 双重注意力机制

本文通过注意力机制指导网络重点关注行人未被遮挡的区域,增加行人关键部位的特征权重,从而避免背景遮挡等干扰信息的影响。针对检测问题中分类和定位两个方面采用不同的注意力机制:在定位支路采用空间注意力机制,在分类支路采用通道注意力机制。同时,利用数据集中提供的行人标签中的全身边界框和可见边界框来为空间注意力机制提供监督信息,从而更加有效地指导网络学习。

1.2.1 空间注意力机制

空间注意力机制的基本思想是通过网络生成一个与原始特征图相同尺寸的掩膜,掩膜中每个元素的值代表特征图对应位置像素的权重,经过学习不断调整各个权重,其本质是告诉网络应该关注的区域。本文的空间注意力机制子网络的结构如图 2 所示。首先通过 4 个大小为 3×3、通道数均为 256 的卷积核对回归分支进行卷积,再利用一个通道数为 1 的 3×3 卷积将特征图压缩成掩膜。为了保留原本的背景信息,以 exp(掩膜参数)乘到原来的特征图上,从而调节原本特征图各个位置的权重。本文为了指导空间注意力机制的学习,使用行人的监督信息生成一个像素级的目标掩膜作为空间注意力机制的标签:将行人的全身边界框和可见边界框区域像素值分别设为 0.8 和 1,其余背景区域像素值设为 0。这样的标签将会指导空间注意力机制关注图片中行人区域,同时更加关注行人的可见区域。

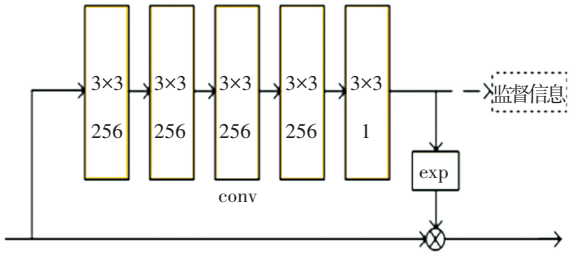


图 2 空间注意力子网络结构

Fig.2 Structure of spatial-wise attention mechanism

1.2.2 通道注意力机制

通道注意力机制基于对卷积神经网络的一个基本认识:卷积特征图的不同通道编码了物体不同部位的特征。文献[16,21]发现一些通道的特征图对行人的特定部位如头、上身和脚等有极高的响应。通道注意力机制的基本思想就是通过网络生成一个长度等于通道数目的向量,向量中的每个元素对应特征图每个通道的权重,通过学习不断调整各通道的权重,其本质是告诉网络应该关注的行人部位。因此本文在分类支路加入通道注意力机制,其网络结构如图 3 所示,与文献[16]中的结构类似:首先对分类支路进行池化;将池化后的权重向量送入全连接层 FC1 和 FC2,对其进行“压缩”和“拉伸”操作;然后通过 sigmoid 函数将向量的分量限制在 0~1 之间,并将两个向量相加融合为最终的权重向量。不同于文献[16]中仅使用平均池化,本文同时采用全局池化和最大池化,这样可以在保留每个通道平均特征的同时突出其主要特征,使得网络更加关注行人的可见部位。

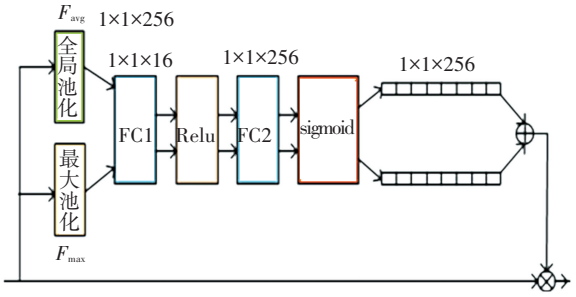


图 3　通道注意力子网络结构

Fig.3　Structure of channel-wise attention mechanism

1.3　损失函数

1.3.1　算法整体损失函数

本文通过一个多任务损失函数联合地对各个部分进行参数调优，该损失函数由 3 个部分组成：

$$\text{Loss} = \frac{1}{M^{c}}\sum_{n\in A} L_{c}(p_n, p_n^{*}) + \lambda_1 \frac{1}{M^{r}} L_{r}(t, t^{*}) + \lambda_2 L_{a}(m, m^{*}) \tag{1}$$

其中：$L_c(p_n, p_n^*)$ 为文献[18] 中设计的改进分类损失函数，基本形式为加权的交叉熵损失函数，其主要目的是改善基于回归的目标检测算法中的正负样本极端不平衡的问题；M^c 为所有预测框的数目；p_n、p_n^* 分别表示预测的第 n 个行人框的类别概率以及相应的实际类别；$L_r(t, t^*)$ 为本文提出的新型回归损失函数，其可以根据不同遮挡程度自主设计权重的大小，下文将具体介绍其设计思路和细节；M^r 为所有预测框的数目，仅考虑判断为前景的部分；$L_a(m, m^*)$ 为空间注意力机制子网络的损失函数，其实际上是一个基于掩膜每个像素的交叉熵损失函数；m, m^* 分别为空间注意力机制生成的掩膜及其对应的掩膜标签；λ_1 和 λ_2 是用来平衡子损失函数的参数，本文中均设为 1。

1.3.2　遮挡感知的回归损失函数

在通用目标检测中，经典的回归损失函数为 smoothL1 函数，其形式为

$$L_{r}(t, t^{*}) = \sum_{n\in A}(p_n^{*} = 1)\sum_{i\in\{x,y,w,h\}} \text{smoothL1}(t_i^{n} - t_i^{*n}) \tag{2}$$

$$\text{smoothL1}(x) = \begin{cases} 0.5x^2, & |x| \leqslant 1 \\ |x| - 0.5, & \text{其他} \end{cases} \tag{3}$$

其中：A 为所有参与计算的行人检测框，t_i^n 为检测的第 n 个行人框，t_i^{*n} 则为其真实坐标，x、y、w、h 分别为真值框的中心点坐标以及宽高。

为了进一步处理遮挡问题，本文提出一种可以依据遮挡程度自主调整检测框权重的回归损失函数。其基本思路是：在计算回归损失函数时，通过预测行人边界框与数据集提供的行人可见区域边界框的 IOG（intersection over ground truth）作为每个正样本产生损失函数的权重，即若预测的正样本边界框与行人可见区域重叠较多，那么它产生的损失更为可信，分配较高的权重，反之则分配较低的权重。基于这个直观的想法，设计出的改进回归损失函数具体形式为

$$L_{r}(t, t^{*}) = \sum_{n\in A}\text{IOG}^{n}(p_n^{*} = 1)\sum_{i\in\{x,y,w,h\}} \text{smoothL1}(t_i^{n} - t_i^{*n}) \tag{4}$$

$$\text{IOG} = \frac{b_{\text{pred}} \cap b_{\text{gtvis}}}{b_{\text{gtvis}}} \tag{5}$$

其中：n 为第 n 个预测框，b_{pred} 为判定为前景的行人预测框，b_{gtvis} 为其对应的行人可见区域边界框。

采用 IOG 而不是 IOU 的原因在于，期望的权重在 0~1 之间，而即使是完全正确的预测框，其与可见区域的 IOU 也可能是一个较小的数值，因此使用 IOG 更为合适。文献[13] 中同样利用行人可见区域与预测框的重叠程度改善遮挡问题，做法是当预测框与行人全身边界框和可见区域边界框的 IOU 同时大于一个固定的阈值时，才判定此预测框为正样本。这种做法有两个不足之处：一是阈值的大小不好设定，二是判定条件过严可能导致有真实框没有对应的预测框。本文提出的新型回归损失函数则有效地解决了这两个问题，更加灵活地利用行人可见框来指导网络的学习。

2　实验结果与分析

2.1　实验设置

2.1.1　数据集

实验是在 Caltech 和 CityPerson 两个行人数据集上进行的。Caltech 数据集[22] 是目前最为常用的公开数据集之一，原始图片为 640×480 像素，提供行人全身边界框和可见区域边界框标签。预先划分好训练集 4 250 张，测试集 4 024 张。CityPerson 是目前较新的公开行人检测数据集，由文献[23] 于 2017 年提供，其包含了 5 000 张德国各地的实拍图片。相比于 Caltech 数据集，其行人遮挡问题更加严重。数据集预先将 2 975 张作为训练集，1 525 张作为测试集，图片为 2 048× 1 024 像素，提供行人全身边界框和可见部分边界框。

2.1.2　先验边界框的设置

本文算法在 5 个不同的特征层进行预测，所以需要设计各特征层上的预设边界框，预设边界框的好坏直接影响到回归的速度与精度。文献[6,9] 手工设计几个固定尺寸和比例的边界框，但其不够灵活且效果稍差。文献[24] 提出了一种更为灵活的

方法,通过对训练集进行聚类来确定预设边界框的尺寸和比例。

本文采用聚类的思想来设计预设边界框,与文献[24]不同的是,本文算法在 5 个特征层进行预测,所以需要根据不同特征图的尺寸合理安排不同大小的预设边界框。具体做法是:首先获得训练集中所有真实框的宽高 $b_{\text{all}}=\{b_1,b_2,\cdots,b_n\}$,为了避免聚类中心被数据量最大的中等尺寸的框主导,预先按框的高度 h 从小到大将所有框划分为 5 份,然后利用 k-means 聚类基于每份边界框生成 9 个预设的边界框,最终共生成 45 个不同大小与比例的预设边界框,分别配置到不同尺度的预测特征层上。聚类中,考虑输出行人边界框的目的,距离度量采用如下形式:

$$d(\text{box},c_i)=1-\text{IOU}(\text{box},c_i) \tag{6}$$

$$\text{IOU}(\text{box},c_i)=\frac{\text{box}\cap c_i}{\text{box}\cup c_i} \tag{7}$$

其中:box 为训练集中的行人边界框,c_i 为第 i 个聚类中心代表的边界框。

2.1.3 训练细节

利用水平翻转、裁剪等操作实现数据增强,增加训练样本集的丰富程度。为了保证图片放缩过程中物体不会变形,通过加 padding 缩放操作将 Caltech 和 CityPerson 数据集的输入图片尺寸分别调整为 1 200×900 像素和 1 400×700 像素,兼顾性能和速度。通过 Adam 算法对网络各部分参数进行优化,学习率的初始值设为 0.000 1,如果连续 3 个 epoch 整体损失函数值不发生明显变化,学习率衰减为原来的 1/10,总训练 epoch 数为 80。主干网络 ResNet50 采用在 ImageNet 上训练好的模型。batch 大小为 2,训练平台为英伟达 RTX 2080。

2.1.4 评估指标

对数平均漏检率(log-average miss rate)[22]是评估行人检测算法最为常用的指标之一。同时为了更好地体现算法对遮挡问题的处理能力,利用数据集提供的行人可见边界框和全身边界框的比值(可见度,Vis)来衡量遮挡程度,将测试集按遮挡程度分为以下 3 类:1)轻微遮挡,Vis>0.65;2)严重遮挡,0.20<Vis<0.65;3)整体,Vis>0.20。分别测试算法在不同遮挡测试集上的检测效果。

2.2 实验结果

本文以 RetinaNet 为基本框架,分别添加双重注意力机制子网络和可感知遮挡的优化回归损失函数,其余参数值均保持和 RetinaNet 相同。消融实验结果见表 2、3。其中 k-means_anchor、attention、weightloss 分别表示是否用聚类预测边界框、是否加入注意力机制子网络以及是否使用改进的回归损失函数。

表 2 Caltech 数据集上消融实验结果

Tab.2 Results of ablation experiments on Caltech dataset

方法	k-means_anchor	attention	weightloss	轻微遮挡性能/%	严重遮挡性能/%	整体性能/%
RetinaNet	×	×	×	13.50	57.72	23.80
改进 1	√	×	×	12.06	56.03	22.27
改进 2	×	√	×	12.23	51.50	21.37
改进 3	×	×	√	12.73	50.37	21.72
本文方法	√	√	√	**9.97**	**45.69**	**18.72**

表 3 CityPerson 数据集上消融实验结果

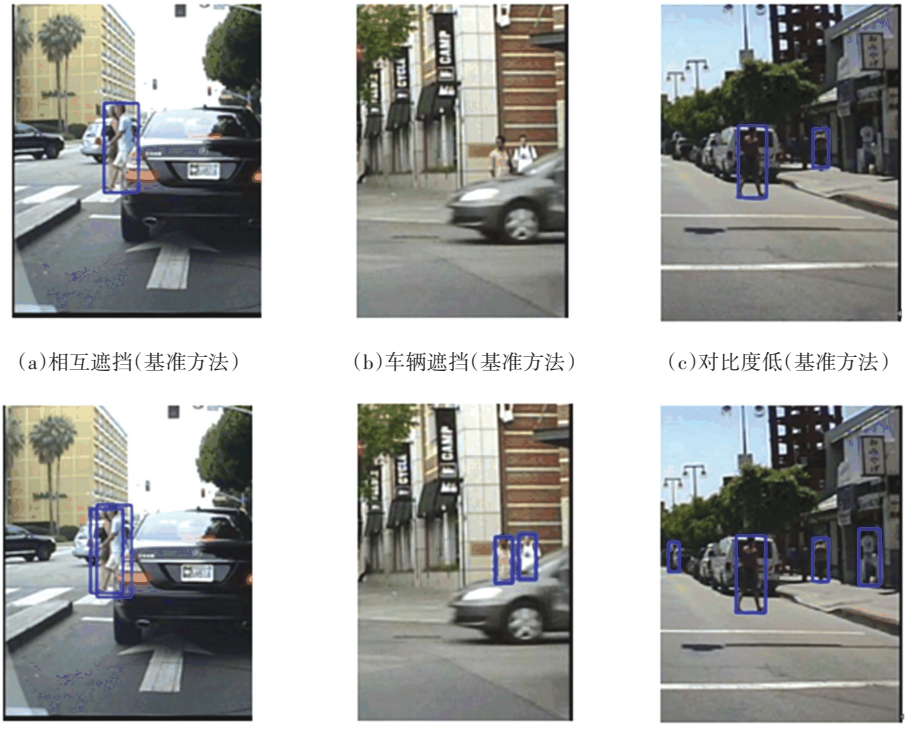
Tab.3 Results of ablation experiments on CityPerson dataset

方法	k-means_anchor	attention	weightloss	轻微遮挡性能/%	严重遮挡性能/%	整体性能/%
RetinaNet	×	×	×	33.63	57.23	48.53
改进 1	√	×	×	31.34	56.03	46.24
改进 2	×	√	×	28.39	53.94	43.02
改进 3	×	×	√	28.48	53.86	43.19
本文方法	√	√	√	**27.43**	**52.67**	**41.95**

由表 2、3 的消融实验结果可以看出,与基准方法 RetinaNet 相比,增加注意力机制子网络和感知遮挡的新型回归损失函数在不同遮挡程度子集上均会带来一定的性能提升,尤其是对于严重遮挡的子集,性能提升更加显著。在 Caltech 的严重遮挡子集上二者分别提高了 6.22%和 7.35%,在 CityPerson 上分别提高了 3.29%和 3.37%。相较于基本框架 RetinaNet,本文方法在 Caltech 和 CityPerson 的严重遮挡子集上分别提高了 12.03%和 4.56%,充分表明该方法对复杂交通场景下的遮挡问题具有很好的处理能力,本文方法

的整体性能与 RetinaNet 相比，同样有较大的提升。此外，利用聚类生成预设行人边界框会给整体性能带来一定提升，但对于遮挡问题效果不明显。

图 4 展示了 Caltech 数据集下基准方法 RetinaNet 与本文方法的检测效果，可以看出，基准方法无法检出一些被汽车、草丛等遮挡的行人，而本文方法可以检出这些目标；对于一些行人之间相互遮挡的现象，基准方法只会给出一个大的边界框，本文方法能分别将每个行人框出，表明本文方法对于类内遮挡和类间遮挡均具有较好的鲁棒性。



(a)相互遮挡(基准方法)　(b)车辆遮挡(基准方法)　(c)对比度低(基准方法)

(d)相互遮挡(本文方法)　(e)车辆遮挡(本文方法)　(f)对比度低(本文方法)

图 4　Caltech 数据集上检测效果图

Fig.4　Detection results of RetinaNet and proposed method on Caltech dataset

在中国的街道场景中，行人更加密集，极易发生遮挡现象，尤其是类内遮挡较为普遍。本文方法对于类内遮挡的鲁棒性结论在国内智能交通领域具有较高的应用价值。

2.3　实验分析

2.3.1　检测效果

表 4 对比了本文方法和其他 8 种方法(包括传统的 HOG[2] 和 ACF[4]，基准 RetinaNet[18]，新近提出的 AdaptFasterRCNN[23] 等)的检测效果。考虑到 CityPerson 数据集较新，目前尚未有充足方法在其上进行测试，且其 Benchmark 并未给出其他方法的原始检测文件，难以客观地与其对比算法性能。因此对比实验只在 Caltech 上进行.

从表 4 中可以看出，本文方法在整体数据集上的平均对数漏检率最低，仅为 18.72%，与其他方法相比具有一定的优势。在轻微遮挡子集上，本文方法的检测效果也处于前列，平均漏检率为 9.97%，略高于 AdaptFastRCNN 等 4 种方法，这可能是本文方法侧重于遮挡问题而导致一些小尺寸行人的漏检。但在严重遮挡子集上，本文方法的性能十分突出，其平均对数漏检率仅为 45.69%，比其他方法小 12%以上，远远领先其他方法。这表明本文针对遮挡问题专门设计的双重注意力机制和遮挡感知的新型回归损失函数非常有效。从表 4 中还可以看出，对于行人检测这样的非刚体、背景复杂且存在遮挡的检测问题，包括本文方法在内的基于卷积神经网络的方法远远好于传统的手工设计特征的方法。

表 4　Caltech 数据集上与其他 8 种方法对比结果

Tab.4　Comparison results of nine methods on Caltech dataset

方法	平均对数漏检率/%		
	轻微遮挡	严重遮挡	整体
HOG[2]	73.34	96.33	78.46
ACF[4]	51.36	94.72	60.91
RPN+BF[25]	9.58	74.36	24.01
SA-FastRCNN[26]	9.68	64.35	21.92
DeepParts[10]	11.89	60.42	22.80
MS-CNN[27]	9.95	59.94	21.53
AdaptFasterRCNN[23]	**9.18**	57.58	20.03
RetinaNet[18]	13.50	57.72	23.80
本文方法	9.97	**45.69**	**18.72**

图 5 进一步给出了严重遮挡子集上几种方法的漏检率随着每张图误检数目变化的曲线图，曲线下

方的面积越小,行人检测算法的性能更强。同样可以看出,随着每图误检数量的变化,本文方法的漏检率都处于最低水平,相比于其他行人检测方法,整体优势明显。

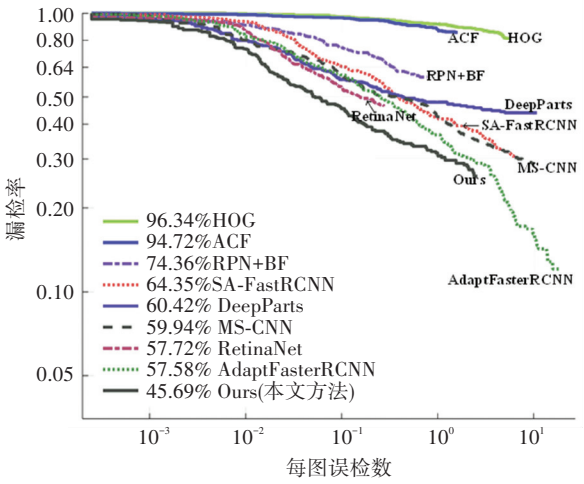


图 5 Caltech 严重遮挡子集上的检测效果对比

Fig.5 Detection results of nine methods on heavy occlusion subset (Caltech dataset)

2.3.2 检测速度

本文方法在 Caltech(缩放至 1 200×900 像素)和 CityPerson 数据集(缩放至 1 400×700 像素)上的检测速度分别为 11.8 帧/s(frames per second)和 10.0 帧/s,实现了准实时的行人检测。此外,也比较了本文方法和其他 4 种精度较高的检测方法(包括 RPN+BF[25], SA-FastRCNN[26]等)的检测效率,实验在 Caltech 数据集上进行。为进行公平的比较,类似于文献[28],本文对比了各方法在单位算力(TFLOPS,每秒万亿次单精度浮点计算)下的检测速度,结果见表 5(GPU 计算能力来自 NVIDIA 官网)。

表 5 5 种方法的检测速度

Tab.5 Detection speed of five methods

方法	设备	TFLOPS	检测速度/(帧·s^{-1}·$TFLOPS^{-1}$)
RPN+BF[25]	Tesla K40	3.5	0.571
SA-FastRCNN[26]	TitanX	5.2	0.325
MS-CNN[27]	TitanX	5.2	0.480
RetinaNet[18]	RTX2080	10.1	1.416
本文方法	RTX2080	10.1	1.168

从表 5 中可以看出,本文方法的检测速度略慢于 RetinaNet,比其他 3 种方法的快 1 倍以上。主要原因在于:本文方法采用单阶段的检测框架,可以实现端到端的快速检测;而 SA-FastRCNN 等 3 种方法采用双阶段的检测框架,需要通过网络生成候选区域然后再进行检测;此外,由于注意力机制子网络带来了附加的计算量,因此本文方法的检测效率略低于 RetinaNet。

3 结 论

提出一种结合双重注意力机制的遮挡感知方法来提高行人检测算法在严重遮挡情况下的性能,降低遮挡对检测造成的影响。该方法通过引入空间/通道双重注意力机制,以及遮挡感知的新型损失函数,能够有效地处理遮挡问题,在 Caltech 和 CityPerson 数据集上分别取得 18.72%和 41.95%的平均漏检率,优于 RetinaNet 等 8 种先进的行人检测算法;尤其在 Caltech 严重遮挡子集上,其平均漏检率仅为 45.69%,低于其他方法 12%以上。并且,该方法可以实现准实时的行人检测,在 Caltech 和 CityPerson 上的检测速度分别为 11.8 帧/s 和 10.0 帧/s。

参考文献

[1] 苏松志,李绍滋,陈淑媛,等.行人检测技术综述[J]. 电子学报, 2012,40(4): 814
SU Songzhi, LI Shaozi, CHEN Shuyuan, et al. Overview of pedestrian detection technology[J]. Electronic Journal, 2012,40(4): 814.DOI:10.3969/j.issn.0372-2112.2012.04.031

[2] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005: 886

[3] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627.DOI:10.1109/TPAMI.2009.167

[4] DOLLAR P, APPEL R, BELONGIE S, et al. Fast feature pyramids for object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(8): 1532. DOI:10.1109/TPAMI.2014.2300479

[5] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. Lake Tahoe: IEEE, 2012: 1097

[6] 姚群力,胡显,雷宏.深度卷积神经网络在目标检测中的研究进展[J].计算机工程与应用, 2018, 54(17): 1
YAO Qunli, HU Xian, LEI Hong. Research of deep convolution neural network in objection detection[J]. Computer Engineering and Application, 2018, 54(17): 1. DOI: 10.3778/j.issn.1002-8331.1806-0377

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]//Advances in Neural Information Processing Systems. Montreal: IEEE, 2015: 91

[8] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779

[9] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector [C]//European Conference on Computer Vision.

Cham：Springer，2016：21

[10]TIAN Y，LUO P，WANG X，et al. Deep learning strong parts for pedestrian detection[C]//IEEE International Conference on Computer Vision. Santiago：IEEE，2015：1904

[11]WANG X，XIAO T，JIANG Y，et al. Repulsion loss：detecting pedestrians in a crowd[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：7774

[12]ZHANG S，WEN L，BIAN X，et al. Occlusion-aware R-CNN：detecting pedestrians in a crowd[C]//European Conference on Computer Vision. Munich：Springer，2018：637

[13]ZHOU C，YUAN J. Bi-box regression for pedestrian detection and occlusion estimation[C]//European Conference on Computer Vision. Munich：Springer，2018：135

[14]徐诚极，王晓峰，杨亚东. Attention-YOLO：引入注意力机制的 YOLO 检测算法[J]. 计算机工程与应用，2019，55(6)：13
XU Chengji，WANG Xiaofeng，YANG Yadong. Attention-YOLO：YOLO with attention mechanism[J]. Computer Engineering and Application，2019，55(6)：13

[15]JADERBERG M，SIMONYAN K，ZISSERMAN A. Spatial transformer networks[C]//Advances in Neural Information Processing Systems. Montreal：IEEE，2015：2017

[16]HU J，SHEN L，SUN G. Squeeze-and-excitation networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：7132

[17]ZHANG S，YANG J，SCHIELE B. Occluded pedestrian detection through guided attention in CNNs[C]//IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：6995

[18]LIN T Y，GOYAL P，GIRSHICK R，et al. Focal loss for dense object detection[C]//IEEE International Conference on Computer Vision. Venice：IEEE，2017：2980

[19]HE K，ZHANG X，REN S，et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE，2016：770

[20]LIN T Y，DOLLAR P，GIRSHICK R，et al. Feature pyramid networks for object detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：2117

[21]GONZALEZ G A，MODOLO D，FERRARI V. Do semantic parts emerge in convolutional neural networks？[J]. International Journal of Computer Vision，2018，126(5)：476. DOI：10.1007/s11263-017-1048-0

[22]DOLLAR P，WOJEK C，SCHIELE B，et al. Pedestrian detection：an evaluation of the state of the art[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2012，34(4)：743. DOI：10.1109/TPAMI.2011.155

[23]ZHANG S，BENENSON R，SCHIELE B. CityPersons：a diverse dataset for pedestrian detection[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：3213

[24]REDMON J，FARHADI A. YOLO9000：better，faster，stronger[C]//IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：7263

[25]ZHANG L，LIN L，LIANG X，et al. Is Faster R-CNN doing well for pedestrian detection？[C]//European Conference on Computer Vision. Cham：Springer，2016：443

[26]LI J，LIANG X，SHEN S M，et al. Scale-aware Fast R-CNN for pedestrian detection[J]. IEEE Transactions on Multimedia，2018，20(4)：985. DOI：10.1109/TMM.2017. 2759508

[27]CAI Z，FAN Q，FERIS R S，et al. A unified multi-scale deep convolutional neural network for fast object detection[C]//European Conference on Computer Vision. Cham：Springer，2016：354

[28]邢浩强，杜志岐，苏波. 基于改进 SSD 的行人检测方法[J].计算机工程，2018，44(11)：228
XING Haoqiang，DU Zhiqi，SU Bo. Pedestrian detection method based on modified SSD[J]. Computer Engineering，2018，44(11)：228. DOI：10.19678/j.issn. 1000-3428.0048553

（编辑　魏希柱）