DOI:10.19651/j.cnki.emt.2106129

基于改进 YOLOv3 的密集行人检测*

邓 杰[1,2] 万旺根[1,2]
(1.上海大学 通信与信息工程学院 上海 200072;2.上海大学 智慧城市研究院 上海 200072)

摘 要:行人检测是目标检测领域的一个重要分支,目前行人检测算法已经取得了较好的发展,但拥挤场景下存在着行人间的严重遮挡,这为检测任务带来了极大地挑战。为有效缓解该问题,在 YOLOv3 的基础上进行改进,提出单阶段密集行人检测算法:Crowd-YOLO,该算法将可见框标注信息加入到网络中,使网络同时预测全身框与可见框信息从而提升检测性能;提出时频域融合注意力模块(TFFAM),将频域通道注意力和空间注意力加入到网络中重新分配特征权重;采用数据关联型上采样代替传统的双线性插值,使深层特征图获取更为丰富的信息表达。使用非常具有挑战性的大型拥挤人群场景数据集 CrowdHuman 进行训练和测试,实验结果表明,所提方法比基础网络在 AP_{50} 指标上提高了约 3.7%,在召回率(Recall)指标上提高了 3.4%,其中时频域融合注意力模块为网络带来了 2.3%AP 的性能增益。实验结果验证了所提方法在拥挤人群场景下的有效性。

关键词:行人检测;遮挡问题;YOLO;融合注意力;数据关联型上采样
中图分类号:TP391.41;TP332 **文献标识码**:A **国家标准学科分类代码**:520.60

Dense pedestrian detection based on improved YOLOv3

Deng Jie[1,2] Wan Wanggen[1,2]
(1. School of Communication and Information Engineering, Shanghai University, Shanghai 200072, China;
2. Institute of Smart City, Shanghai University, Shanghai 200072, China)

Abstract: Pedestrian detection is an important branch in the field of object detection, and pedestrian detection algorithms have been well developed, but there exists severe occlusion between pedestrians in crowded scenes, which makes a great challenge for the detection task. To effectively alleviate this problem, this paper improves on YOLOv3 and proposes a single-stage dense pedestrian detection algorithm: Crowd-YOLO, which adds visible frame labeling information to the network to assist training, so that the network can predict both full-body frame and visible frame information to improve detection performance, proposes a time-frequency domain fused attention module (TFFAM), which adds frequency-domain channel attention and spatial attention to the network to redistribute features, uses data correlation upsampling instead of traditional bilinear interpolation to obtain a richer information representation of deep feature maps. This paper uses the very challenging large crowd scenario dataset CrowdHuman for training and testing. The experimental results show that the proposed method improves the AP_{50} metric by about 3.7% and the recall metric by 3.4% over the baseline, with the time-frequency domain fused attention module bringing a 2.3% AP performance gain. The experimental results verify the effectiveness of the proposed method in crowded scenarios.

Keywords: pedestrian detection; occlusion problem; YOLO; fused attention; data correlation upsampling

0 引 言

行人检测是目标检测算法研究中的重要研究方向,也是计算机视觉领域中一个极具挑战性的课题。行人检测的目的是检测出图片或视频序列中是否存在行人并标记其位置,应用于智能安防、智能机器人以及自动驾驶等领域。行人检测常常作为行人跟踪、行人重识别以及行人搜索等任务的重要前置处理环节,行人检测算法的性能会密切影响后续任务的效果,因此提升行人检测算法的精度有非常重要的意义。

传统行人检测方法使用手工设计的特征来训练分类器,用于区分行人和背景。这些特征包括纹理、颜色特征、

收稿日期:2021-03-24
* 基金项目:上海市科委港澳台科技合作项目(18510760300)、中国博士后基金项目(2020M681264)资助

梯度方向直方图、局部二值模式、积分通道特征等。手工设计的特征往往带有设计者的主观意识,另外由于行人目标是非刚性目标,姿态和形状各不相同且衣服颜色、纹理特征多种多样,这些因素都对行人检测带来了很大地挑战。

随着卷积神经网络在计算机视觉领域大放异彩,基于深度学习的目标检测算法也有了很大地突破,行人检测也开始采用深度学习方法。Zhang 等[1]将两阶段目标检测算法 Faster R-CNN[2]应用于行人检测,使用级联的增强森林算法对区域提议网络输出的候选目标区域进行分类。Wang 等[3]提出排斥损失,以应对密集行人场景下行人目标之间的遮挡问题,该损失函数使预测框和真实目标框之间的距离缩小,同时使其与周围非目标框的距离加大。同样为了缓解遮挡问题,Zhang 等[4]提出聚合损失,该方法使检测结果中属于同一个真实目标的预测框尽可能靠近并最小化与同一目标相关联的提议的内部距离,同时提出部分遮挡感兴趣区域池化单元将人体可见部分的先验结构信息整合到网络中来处理遮挡问题。由于单阶段目标检测器的检测速度相比两阶段目标检测器更快,Liu 等[5]在 SSD 的基础上提出了 ALFNet,通过加入渐进定位拟合模块提升了单阶段目标检测网络的准确率,通过叠加一系列预测因子将 SSD 的默认锚框逐渐训练成更好的检测结果。随着无锚框目标检测算法在一般目标检测任务中取得了很大的突破,近年来,行人检测领域也出现了一些无锚框的行人检测算法。Liu 等[6]提出 CSP(center and scale prediction)行人检测器,通过卷积将行人检测简化为中心和尺度预测任务,网络预测行人目标的中心点位置和偏移量以及对应的行人高度值,然后根据行人目标的长宽比先验知识获取宽度信息,生成目标包围框。尽管目前行人检测领域已经取得了重大进展,但是在拥挤人群场景中的行人检测仍然分非常具有挑战性。行人之间的严重遮挡给非极大值抑制(NMS)算法带来了巨大挑战。针对此类问题,Huang 等[7]提出代表性区域非极大值抑制算法,该算法利用较少遮挡的可见部分进行后处理,有效抑制了冗余的检测框并减少了假正例,提出成对盒模型以同时预测全身框与可见框。Wang 等[8]提出一种改进的多属性行人检测方法,通过考虑密度信息和类平衡策略来降低小目标的假阳性,并结合 ID 信息提出多属性 NMS 算法,利用更多的行人属性信息来重新定义检测结果。

为有效缓解行人检测领域的遮挡问题,本文提出单阶段密集行人检测算法:Crowd-YOLO,使用非常具有挑战性的 CrowdHuman[9]数据集进行训练和测试。本文的主要工作如下:1)使用 K-means 算法对 CrowdHuman 数据集中的行人尺寸进行聚类,预定义锚框将加速网络收敛;2)为利用该数据集丰富的标注信息以及减少漏检,本文将遮挡区域较少的行人可见框信息加入到网络中,在损失函数中为可见框信息分配较低权重来辅助网络训练;3)在网络输出阶段的卷积层后加入时频域融合注意力模块(time and frequency domain fusion attention module,TFFAM),在频域通道及时域空间层面为特征图分配合理权重,使网络更好地关注行人区域;4)改进上采样方式,使用数据关联型上采样[10]代替传统的双线性插值,可以取得更好的高分辨率特征图并且保留更为丰富的深层特征信息。为了使检测网络达到精度与速度的权衡,并验证以上方法的有效性,本文将采用一阶段目标检测算法 YOLOv3[11]作为基础网络。

1　CrowdHuman 数据集

CrowdHuman 数据集主要针对拥挤人群检测任务,考虑了大量的拥挤人群场景。该数据集规模庞大,训练集、验证集和测试集分别有 15 000、44 370、5 000 张图片,在训练集和验证集中大约包含 47 万人,每张图片的行人数量达到了 22.6,同时存在着各种各样的遮挡情况。数据集包含丰富的标注信息,每个人类实例都标有头部、可见区域和全身区域的包围框。相比 Caltech[12]、KITTI[13]、CityPersons[14]、COCOPersons[15]等行人检测数据集,CrowdHuman 数据集拥有更为丰富的信息标注以及更高的人群密度和更为频繁的遮挡情况。表 1 所示为 CrowdHuman 数据集与其他行人检测数据集在信息标注方面的对比,表 2 所示为 CrowdHuman 数据集与其他行人检测数据集在图片数据量、标注实例数目及人群密度(每张图片包含的平均人数)等方面的对比。

表 1　不同行人检测数据集标注信息对比

标注信息	Caltech	KITTI	CityPersons	COCOPersons	CrowdHuman
全身框	√	√	√	×	√
可见框	√	×	√	√	√
头部框	×	×	×	×	√

表 2　不同行人检测数据集详细对比

对比内容	Caltech	KITTI	CityPersons	COCOPersons	CrowdHuman
图片数量	42 782	3 712	2 975	64 115	15 000
人类实例数量	13 674	2 322	19 238	257 252	339 565
忽略区域数量	50 363	45	6 768	5 206	99 227
人群密度	0.32	0.63	6.47	4.01	22.64

2　研究方法

2.1　K-meams 聚类预定义锚框

YOLOv3 是基于锚框的单阶段目标检测算法，对于网络输出特征图的每一个网格都会施加 3 个不同长宽比的锚框，用于对多尺度物体的预测[16]。

本文中仍然使用 K-meams 算法对 CrowdHuman 数据集中的全身框与可见框尺寸进行聚类，并对特征图的每个网格施加 4 个不同尺寸的锚框，其中 2 个属于全身框，2 个属于可见框。重新聚类后的锚框可以使网络更快地适应数据集中行人的尺寸变化，加速网络的收敛。K-means 聚类后的锚框尺寸如表 3 所示。

表 3　K-meams 聚类 CrowdHuman 数据集预定义锚框

特征图尺寸	13×13	26×26	52×52
感受野	大	中	小
全身区域锚框	(105×64)(144×435)	(35×138)(65×251)	(8×22)(18×62)
可见区域锚框	(100×82)(127×298)	(31×115)(55×217)	(7×18)(18×50)

由于 CrowdHuman 数据集中的类别只有 person 和 mask 两类(mask 为类人物体，比如人形雕塑)，而网络最终会输出全身框与可见框的预测信息，包括中心点与宽高的偏移量和置信度得分，因此网络输出特征图的维度变为 48 维。CrowdHuman 数据集中全身框与可见框标注的示例如图 1 所示，其中实线框为全身框标注，虚线框为可见框标注。

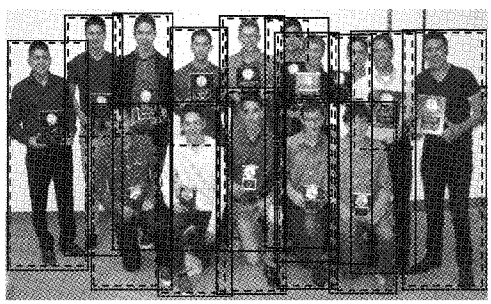

图 1　CrowdHuman 中全身框与可见框的标注

2.2　加入可见框标注信息辅助训练

CrowdHuman 数据集中对每个人类实例都标注有全身框、可见框与头部框信息。该数据集人群密度较高，目标之间有较大的重叠度，若仅使用全身框标注信息进行训练，不仅将带来目标定位的困难，同时在进行非极大值抑制(NMS)后处理算法时会抑制掉更多的真正例，造成行人目标的漏检。本文认为全身框与可见框在空间与特征上存在着紧密的联系，可见框的加入意味着网络将会重点学习两者之间的重叠部分，这部分特征将有助于网络更准确地定位目标。因此本文将可见框信息加入到网络中，在设计损失函数时为可见框信息分配较小的权重来辅助训练，实验结果表明在仅使用原始 NMS 的情况下仍然带来了平均精度(AP)的提升。加入可见框信息后的损失函数如下所示：

$$loss = loss_{cls} + loss_{location} + loss_{conf} \tag{1}$$

式中：$loss_{cls}$ 为分类损失，使用二元交叉熵损失函数；$loss_{location}$ 为定位损失，包括全身框与可见框的中心点与宽高偏移量损失，使用均方损失函数；$loss_{conf}$ 为边框置信度损失，使用二元交叉熵损失函数。$loss_{location}$ 定义如下：

$$loss_{location} = \alpha loss_{fullbody} + \beta loss_{visiblebody} \tag{2}$$

式中：α，β 为超参数，实验中 $\alpha=0.8$，$\beta=0.2$。

2.3　时频域融合注意力

目前注意力机制在计算机视觉领域已经取得了巨大的成功。通道注意力是从特征图通道之间的关系入手，显式地建模特征通道之间的相互依赖关系，以网络学习的方式自动获取每一个特征通道的重要程度，从而使网络更加关注于图片中感兴趣的部分。通道注意力权重得分计算如下所示：

$$chan\_att = sigmoid(fc(gap(X))) \tag{3}$$

输入 X 经过全局平均池化处理后输入到全连接层然后经过 sigmoid 激活函数得到加权得分 $chan\_att$，之后 $chan\_att$ 再与原始输入进行逐通道相乘得到注意力特征输出，如下所示：

$$\tilde{X}_i = chan\_att_i \cdot X_i,\ \text{s.t. } i \in \{0,1,\cdots,C-1\} \tag{4}$$

空间注意力则是对目标的空间位置关系进行建模，使网络更准确地获取目标位置。

Qin 等[17]认为在通道注意力机制中，全局平均池化(GAP)是数据预处理的默认方式，而对空间中的所有元素取均值并不足以表达各通道之间的相互关系，因此无法捕获丰富的输入表示。Qin 等从频域角度出发重新思考通道注意力，用数学方法证明了全局平均池化是离散余弦变换(DCT)的最低频分量，该方法首先对输入特征进行分块，然后使用离散余弦变换将分块特征变换为有限制的多个频率分量，来代替只有最低频分量的全局平均池化，从而提出多谱通道注意力机制，该方法采用更多的频率分量从而引入了更多的信息。频域通道注意力模块(FCAM)如图 2 所示。

首先输入特征 X 被划分为多个通道，每个通道都会被分配 1 个二维离散余弦变换分量 $Freq^i$，然后连接所有频率分量得到多谱向量 $Freq$，再将多谱向量输入到全连接层中学习注意力得分，如式(5)、(6)所示。

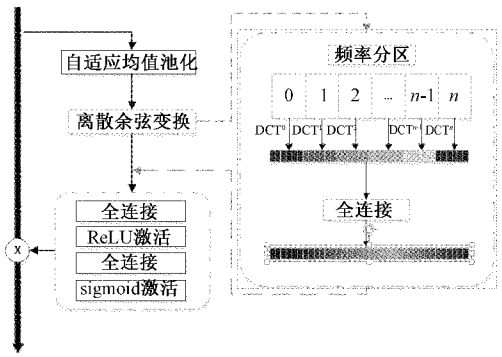

图 2　频域通道注意力模块(FCAM)

$$Freq = cat([Freq^0, Freq^1, \cdots, Freq^{n-1}]) \tag{5}$$

$$fca_{att} = sigmoid(fc(Freq)) \tag{6}$$

本文受 Qin 等[17]和 Woo 等[18]启发,将频域通道注意力模块(FCAM)与空间注意力模块(SAM)相结合,提出时频域融合注意力模块(TFFAM),将其加入到网络输出层的每一个卷积层之后,使网络从频域通道及时域空间层面重新分配深层特征图的权重,使网络更好地关注行人区域。

2.4　数据关联型上采样

YOLOv3 采用特征金字塔[19]网络输出 3 个不同尺寸的特征图来应对多尺度目标检测问题,在每一层特征图输出之前会进行卷积和上采样操作得到高分辨率的深层特征图,并与浅层特征进行融合,浅层特征可以保留目标的位置信息,而深层特征包含了目标更多的抽象特征信息[20]。在深层网络,特征图尺寸已变为原图的 1/16 或者 1/32,此时仅使用双线性插值进行上采样会给特征图带来更多的噪声,同时特征图各通道之间是数据不相关的,这将为网络学习下一输出层的融合特征带来困难。

本文使用数据关联型上采样代替传统的双线性插值,为深浅层特征融合带来更为丰富的特征表达。以 YOLOv3 特征金字塔第 1 层输出特征图的上采样过程为例,数据关联型上采样的网络结构如图 3 所示。给定输入特征图尺寸为 $C\times H\times W$,上采样倍数为 2。对于特征图中每个维度为 $1\times C$ 的像素,通过 1×1 卷积将其维度扩大为 $1\times N$ ($N=4\times C$),然后通过两次置换及重塑操作将 $1\times N$ 的表示转换为 $2\times 2\times N/4$ 的表示,输出的特征图维度变为 $C\times 2H\times 2W$,以此来完成上采样操作。数据关联型上采样考虑了像素点在各通道间与空间上的关系,上采样权重由网络学习得到,可以取得更好的高分辨率特征图并且保留更为丰富的深层网络信息。

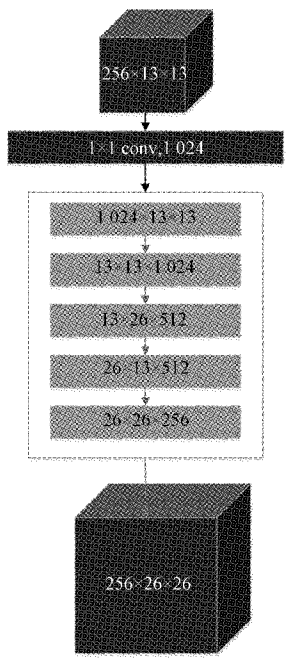

图 3　数据关联型上采样示意图

本文所提方法 Crowd-YOLO 的整体网络结构如图 4 所示。

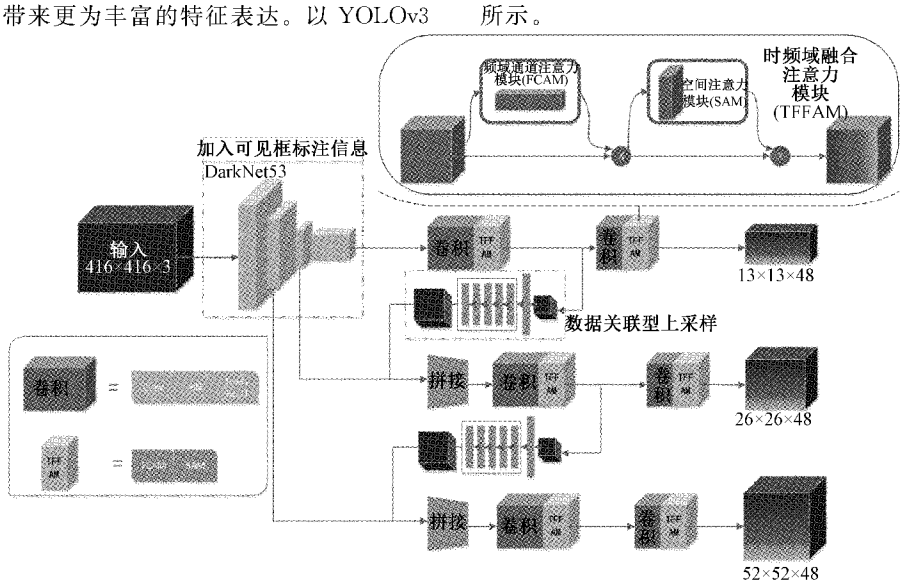

图 4　Crowd-YOLO 网络结构

3　实验分析

3.1　实验设置

本节实验环境配置：Ubuntu16.04（LTS）、GTX 1080Ti 12 GB、CUDA 9.0、Python 3.7、PyCharm 2019.2、PyTorch 1.2.0、torchvision 0.4.0。本文以 YOLOv3 作为基础网络，骨架网络为 DarkNet53，使用 ImageNet 预训练参数。采用随机梯度下降法（SGD）训练 100 个 epoch，batch size 设置为 8，学习率设置为 0.001，NMS 阈值设置为 0.5，输入图片尺寸设置为 416×416，仅使用水平翻转进行数据增强，采用多尺度训练，使用平均精度 AP_{50}（IOU=0.5）、召回率（Recall）及精度（Precision）作为评价指标。

3.2　实验结果分析

为验证本文所提方法在密集行人场景下的有效性，实验使用非常具有挑战性的拥挤人群检测数据集 CrowdHuman 进行训练，并在其验证集上进行测试。表 4 所示为本文方法的实验结果，第 1 行为基准网络 YOLOv3 在 CrowdHuman 数据集上训练后的检测结果，第 2 行为本文所提方法 Crowd-YOLO 在该数据集上的检测结果，在使用原始非极大值抑制（NMS）后处理算法的基础上，本文所提方法较基准网络提高 AP_{50} 约为 3.7%，召回率提高约 3.4%，证明了本文所提方法在拥挤人群场景下的有效性。

表 4　Crowd-YOLO 算法在 CrowdHuman 数据集上的实验结果

方法	AP_{50}	Recall	Precision
YOLOv3(baseline)	33.0	69.5	**36.3**
Crowd-YOLO(本文方法)	**36.7**	**72.9**	36.2

3.3　消融实验分析

表 5 所示为本文所提方法中各个模块的消融实验结果。由表 5 可知，仅加入可见框信息辅助网络训练可提升 AP_{50} 约 1.1%，加入频域通道注意力（FCAM）模块可提升 AP_{50} 约 1.8%，加入空间注意力（SAM）模块可提升 AP_{50} 约 0.5%，而同时加入频域通道注意力和空间注意力，也就是本文第 2.3 节提出的时频域融合注意力（TFFAM）模块可提升 AP_{50} 约 2.3%，最后将双线性插值替换为数据关联型上采样可提升 AP_{50} 约 0.3%。

表 5　Crowd-YOLO 算法各模块在 CrowdHuman 数据集上的消融实验结果

baseline	可见框	FCAM	SAM	上采样	AP_{50}	ΔAP_{50}
√					33.0	—
√	√				34.1	1.1
√	√	√			35.9	1.8
√	√	√	√		36.4	0.5
√	√	√	√	√	36.7	0.3

由实验可知本文方法提出的各个模块都对网络性能起到了促进作用，其中时频域融合注意力模块对网络的提升作用最大，同时注意到频域通道注意力的作用大于空间注意力，这说明在网络训练过程中特征图经过频域划分、离散余弦变换等操作引入多谱频域向量的方法为特征通道重新分配权重对网络带来了最大的收益，而在通道注意力之后加入空间注意力可帮助网络更精确地定位行人区域；其次可见框标注信息的加入使网络可同时预测全身框与可见框，由于两者在空间及特征层面都存在着很强的相关性，重叠部分将成为网络重点学习的对象，即使原始非极大值抑制后处理算法会抑制掉 IOU 较大的检测框，该方法仍然可以保留更多的真正例；最后，数据关联型上采样带来的增益最小，可能是由于在上采样阶段 1×1 的卷积权重是随机初始化得到，导致这部分参数网络并未充分学习，即便如此，数据关联型上采样也为网络性能带来了小幅的提升。

4　结　　论

本文提出了一种面向拥挤人群场景的密集行人检测算法：Crowd-YOLO，该算法从添加可见框标注信息、时频域融合注意力、数据关联型上采样等 3 个方面对原始 YOLOv3 进行改进。可见框与全身框之间存在着天然的强相关性，在训练过程中可辅助网络更好地定位行人区域；时频域融合注意力模块包括频域通道注意力及空间注意力，其中频域通道注意力机制在提升网络性能方面起到了关键作用；数据关联型上采样考虑特征通道间的相关性，为网络性能带来了微小的增益。该算法使用非常具有挑战性的大型拥挤人群数据集 CrowdHuman 进行训练和测试，在仅使用原始非极大值抑制后处理的情况下仍然为检测网络带来了性能的提高，最后的实验结果验证了本文方法在密集人群场景下的有效性。目前目标检测在计算机视觉领域已取得较好的发展，后续研究考虑使用当前检测精度更好的算法进行实验。同时基于锚框的算法也为网络性能带来了负担，特别是在拥挤人群场景下，网络将会生成数量更多的锚框，这意味着正负样本的比例将进一步失衡，如何处理正负样本不均衡的问题也是下一步研究的课题之一。

参考文献

[1] ZHANG L L, LIN L, LIANG X D, et al. Is faster R-CNN doing well for pedestrian detection? [C]. ECCV(2), 2016: 443-457.

[2] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Trans. Pattern Anal. Mach. Intell, 2017, 39(6): 1137-1149.

[3] WANG X L, XIAO T, JIANG Y, et al. Repulsion loss: Detecting pedestrians in a crowd[C]. CVPR,

2018:7774-7783.

[4] ZHANG S F, WEN L Y, BIAN X, et al. Occlusion-aware R-CNN: Detecting pedestrians in a crowd[C]. ECCV(3), 2018: 657-674.

[5] LIU W, LIAO S C, HU W D, et al. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting[C]. ECCV(14), 2018: 643-659.

[6] LIU W, LIAO S C, REN W Q, et al. High-level semantic feature detection: A new perspective for pedestrian detection[C]. CVPR, 2019: 5187-5196.

[7] HUANG X, GE Z, JIE Z Q, et al. NMS by representative region: Towards crowded pedestrian detection by proposal pairing [C]. CVPR, 2020: 10747-10756.

[8] WANG Y, HAN C, YAO G L, et al. MAPD: An improved multi-attribute pedestrian detection in a crowd[J]. Neurocomputing, 2021, 432: 101-110.

[9] SHAO S, ZHAO Z J, LI B X, et al. CrowdHuman: A benchmark for detecting human in a crowd[J]. ArXiv Preprint, 2018, ArXiv: 1805.00123.

[10] TIAN Z, HE T, SHEN C H, et al. Decoders matter for semantic segmentation: Data-dependent decoding enables flexible feature aggregation[C]. CVPR, 2019: 3126-3135.

[11] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv Preprint, 2018, ArXiv: 1804.02767.

[12] DOLLÁR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: A benchmark [C]. CVPR, 2009: 304-311.

[13] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]. CVPR, 2012: 3354-3361.

[14] ZHANG S S, BENENSON R, SCHIELE B. CityPersons: A diverse dataset for pedestrian detection [C]. CVPR, 2017: 4457-4465.

[15] LIN T Y, MAIRE M, SERGE J. et al. Microsoft COCO: Common objects in context[C]. ECCV(5), 2014: 740-755.

[16] 曹红燕,沈小林,刘长明,等. 改进 YOLOv3 的红外目标检测算法[J]. 电子测量与仪器学报, 2020, 34(8): 188-194.

[17] QIN Z Q, ZHANG P Y, WU F, et al. FcaNet: Frequency channel attention networks [J]. ArXiv Preprint, 2020, ArXiv: 2012.11879.

[18] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module[C]. ECCV(7), 2018: 3-19.

[19] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Belongie: Feature pyramidnetworks for object detection[C]. CVPR, 2017: 936-944.

[20] 韩航迪,徐亦睿,孙博,等. 基于改进 Tiny-YOLOv3 网络的航天电子焊点缺陷主动红外检测研究[J]. 仪器仪表学报, 2020, 41(11): 42-49.

作者简介

邓杰,硕士,主要研究方向为计算机视觉、行人检测与重识别。

E-mail: 13262218151@163.com

万旺根,教授,博士生导师,主要研究方向为计算机图形学、信号处理和数据挖掘。

E-mail: wanwg@staff.shu.edu.cn