DOI:10.16515/j.cnki.32-1722/n.2021.02.006

基于SSD多模态多尺度特征融合的行人检测算法

陈 敏[1],王池社[1,2*],郝达慧[1]

(1.安徽理工大学计算机科学与工程学院,安徽 淮南 232001;
2.金陵科技学院网络与通信工程学院,江苏 南京 211169)

摘 要:针对行人检测算法在复杂光照场景以及多尺度情况下检测效果较差的问题,提出了一种基于SSD多模态多尺度特征融合的行人检测算法CMMSNet。该算法使用ResNet作为SSD算法的主干特征提取网络,分别提取可见光图像和红外图像两种模态的特征,在多尺度的特征层上使用注意力机制对两种模态的特征进行加权融合,使用多尺度特征融合层进行行人检测。在KAIST数据集上的实验结果表明,CMMSNet在复杂光照场景以及多尺度情况下具有良好的鲁棒性,行人检测效果显著提升。

关键词:行人检测;SSD;多模态;多尺度;特征融合
中图分类号:TP391 **文献标识码**:A **文章编号**:1672-755X(2021)02-0033-06

Pedestrian Detection Algorithm Based on SSD Multi-modal and Multi-scale Feature Fusion

CHEN Min[1], WANG Chi-she[1,2*], HAO Da-hui[1]

(1. Anhui University of Science and Technology, Huainan 232001, China;
2. Jinling Institute of Technology, Nanjing 211169, China)

Abstract: Aiming at the problem of poor detection effect of pedestrian detection algorithms in complex lighting scenes and multi-scale conditions, a pedestrian detection algorithm CMMSNet based on SSD multi-modal and multi-scale feature fusion is proposed. The algorithm uses ResNet as the backbone feature extraction network in the SSD algorithm to extract the features of the visible light image and the infrared image respectively, and uses the attention mechanism on the multi-scale feature layer to weight and fuse the features of the two modalities. The multi-scale feature fusion layer is used to carry out pedestrian detection. Experimental results on KAIST dataset show that CMMSNet has good robustness in complex lighting scenes and multi-scale situations, and the pedestrian detection effect has been significantly improved.

Key words: pedestrian detection; SSD; multi-modal; multi-scale; feature fusion

行人检测在智慧交通、视频监控、机器人等领域具有广泛的应用,一直是计算机视觉领域最热门的研究方向之一,受到学术界和工业界广泛关注[1-4]。随着Faster R-CNN[5]、SSD[6]及YOLO[7]等相关深度学习算法的引入,行人检测技术取得了突破。目前如何提高复杂光照场景下的行人检测性能仍是一大难点

收稿日期:2021-03-19
基金项目:交通运输部2020年重点项目(2020-ZD-029)
作者简介:陈敏(1995—),男,江苏淮安人,硕士研究生,主要从事人工智能和计算机视觉研究。
通信作者:王池社(1974—),男,安徽铜陵人,教授,博士,主要从事大数据技术、人工智能、机器学习研究。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

问题。可见光模态的图像受光照条件的影响较大。白天光照条件较好时可见光模态图像较为清晰，行人检测效果较好；晚上或者光线阴暗情况时可见光模态图像质量较差，行人检测效果很不理想。红外模态的图像反映人体散发出的热辐射，白天光照强烈会导致红外模态图片背景产生噪声干扰，行人红外图像通常在夜晚或者光线阴暗处表现出明显的轮廓信息，所以红外模态的图像更适合低光场景下的行人检测。

近年来，行人检测方法主要分为两种：基于机器学习手工设计特征的行人检测方法和基于深度学习的行人检测方法。随着卷积神经网络在计算机视觉领域的广泛应用，越来越多的研究人员将研究重心转向基于深度学习的行人检测方法。大部分行人检测算法是基于可见光模态进行研究的，但是在复杂光照条件下，如光线较暗的夜晚，基于可见光模态的行人检测效果较差，因此许多学者引入多模态数据来训练行人检测模型[8-9]，以在复杂光照场景中也能有效地检测行人。Hwang 等[10] 于 2015 年采集并标注了多模态行人检测公开数据集 KAIST(Korea advanced institute of science and technology)，引入了多光谱 ACF (aggregate channel feature)，使用 ACF＋T＋THOG (ACF＋thermal＋thermal histogram of oriented gradients)同时处理可见光-红外图像对，通过 AdaBoost 方法进行目标分类得到行人检测结果。童靖然等[11] 采用锐化融合的方式构建了 FFPN(fusion feature pyramid networks)，将可见光模态和红外模态的强特征进行强化突出，既保留了单模态的强特征，又互补了双模态的弱特征。Liu 等[12] 设计了 Halfway Fusion 模型，在 Faster R-CNN 算法的基础上，采用级联融合方式将多模态特征进行融合，之后放入卷积网络中进行目标检测。Zhang 等[13] 提出一种跨模态交互式注意力网络(cross-modality interactive attention network，CIAN)，首先利用 RGB-热信息流数据建立 2 个分离的特征层次，接着提取全局特征，最后对中间特征映射的通道进行自适应校正。以上方法只在单一特征层上面进行特征融合，本文提出基于 SSD 多模态多尺度特征融合的行人检测算法 CMMSNet，通过多尺度对两种模态的特征进行融合，以提高行人检测的性能。

1 基于 SSD 多模态多尺度特征融合的行人检测算法

1.1 算法整体结构

本文提出的基于 SSD 多模态多尺度特征融合的行人检测算法 CMMSNet 的网络结构如图 1 所示。该算法首先使用 ResNet 替换 SSD 算法中主干特征来提取网络，分别构建可见光和红外两种模态的多尺度特征层，然后使用注意力机制对两种模态构建好的多尺度特征层进行加权融合，最后使用多尺度的特征融合层进行行人检测的分类与位置回归，构建出行人检测器。

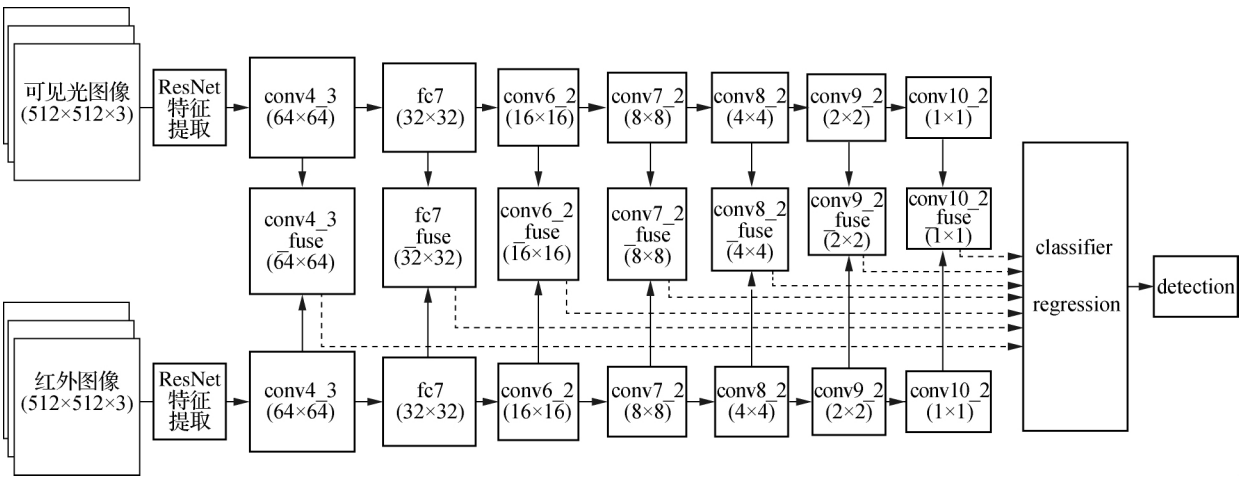

图 1 基于 SSD 多模态多尺度特征融合的行人检测算法

算法训练过程中，先验框(default box)生成规则为：在多尺度特征层上，以每个点为中心生成一系列先验框。使用 m 个尺度特征层生成先验框时，每一层先验框基础边长与原图边长的比例为 S_i，最浅特征层上先验框的比例为 $S_{\min}$，最深特征层上先验框的比例为 $S_{\max}$，其余各层先验框对应比例的计算公式[6] 如下：

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

$$S_i=S_{\min}+\frac{S_{\max}-S_{\min}}{m-1}(i-1),\quad i\in[1,m] \tag{1}$$

其中，m 为多尺度特征层的层数，i 为当前第几个特征层且 i 为整数，训练过程中 $S_{\min}$ 设置为 0.2，$S_{\max}$ 设置为 0.9。由于图像像素点个数都是整数，通过原图尺寸与 S_i 相乘得到的先验框基础边长可能会是小数，导致无法匹配到像素点，因此需要在比例 S_i 的基础上进行调整，调整之后每一层先验框基础边长的计算公式如下：

$$l=\lfloor w\times S_i \rfloor \tag{2}$$

其中 w 为原图的边长，本文算法采用的是 SSD512，即 w 为 512。

算法训练过程中，损失函数(loss function)按照如下公式进行计算：

$$Loss=\frac{1}{N}(L_{\text{conf}}+\partial L_{\text{loc}}) \tag{3}$$

算法训练优化的损失函数由分类损失(L_{conf})和位置回归损失(L_{loc})两部分组成。其中 N 是先验框与标注框匹配的数量，∂参数用于调整分类损失和位置回归损失之间的比例，默认值为 1。

1.2　基于 SSD 生成多尺度特征层

深度学习需要较深的网络层来提取不同模态间的特征，当网络层较深时会出现梯度消失/爆炸的问题，导致网络训练出现退化。ResNet 提供了一种残差结构来解决网络层数较深时网络训练出现退化的问题。本文算法使用 ResNet 替换 SSD 算法中主干特征来提取网络，分别提取可见光和红外模态的特征信息，生成多尺度特征层 conv4_3，fc7，conv6_2，conv7_2，conv8_2，conv9_2，conv10_2，如图 2 所示。

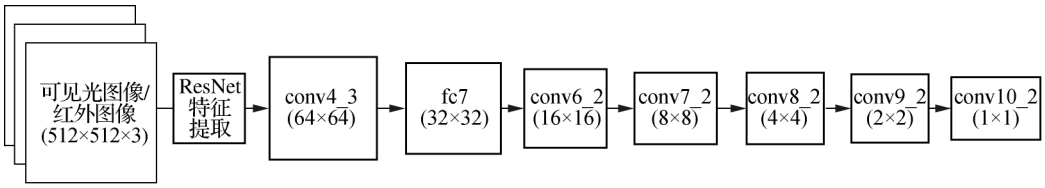

图 2　基于 SSD 生成多尺度特征层

1.3　多尺度特征融合

本文算法分别提取可见光和红外模态的特征信息，生成对应模态的多尺度特征层。在对应模态的多尺度特征层上进行融合，构建多尺度特征融合层 conv4_3_fuse，fc7_fuse，conv6_2_fuse，conv7_2_fuse，conv8_2_fuse，conv9_2_fuse，conv10_2_fuse，如图 3 所示。

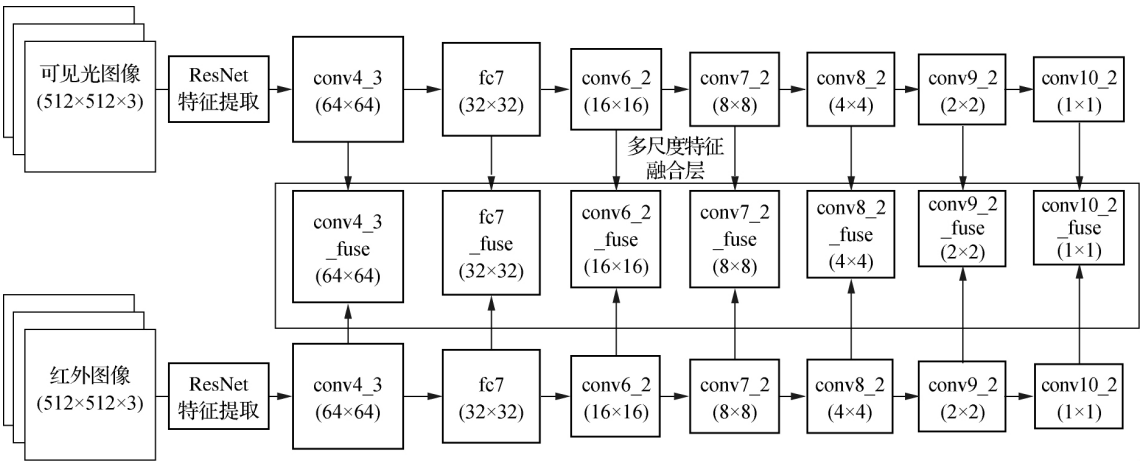

图 3　多尺度特征层融合

在多尺度特征层上使用注意力机制对两种模态的特征进行加权融合，特征层融合模块如图 4 所示，其中 F_c、F_t 为可见光和红外的特征层(feature map)，W_c、W_t 分别为 F_c、F_t 特征层经过全局均值池化之后的特

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

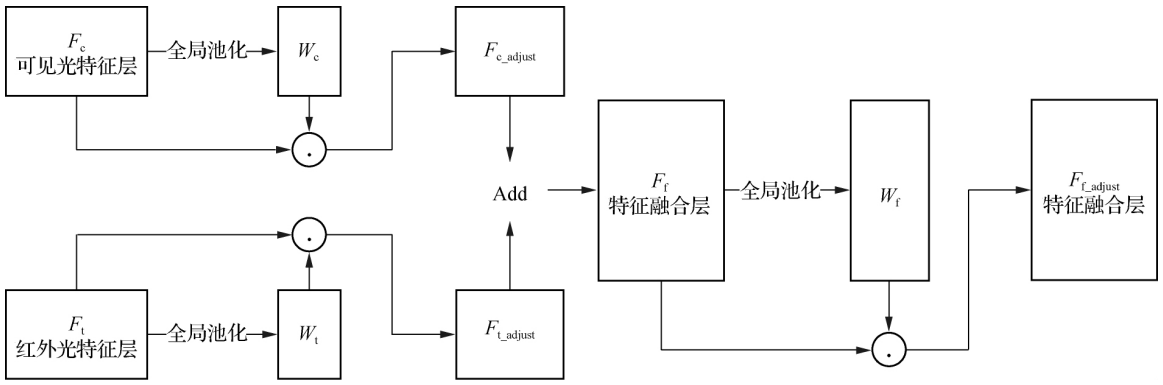

图 4 特征层融合模块

征层的权值，用来调整可见光与红外特征层融合的权重，计算公式如下：

$$W_c = \frac{1}{a \times b}\sum_{j=1}^{b}\left(\sum_{i=1}^{a} F_c^{ij}\right), \quad W_t = \frac{1}{a \times b}\sum_{j=1}^{b}\left(\sum_{i=1}^{a} F_t^{ij}\right) \tag{4}$$

其中，a、b 为特征层的宽高，F_c^{ij}、F_t^{ij} 分别为可见光和红外特征层第 i 行第 j 列的值。

F_{c_adjust}、F_{t_adjust} 分别为可见光和红外特征层权值调整之后的特征层，计算公式如下：

$$F_{c_adjust} = F_c \cdot W_c, \quad F_{t_adjust} = F_t \cdot W_t \tag{5}$$

F_f 为特征融合层，$F_f = F_{c_adjust} + F_{t_adjust}$。$W_f$ 求法同式(4)。

最终，可见光和红外特征融合层完整的计算公式如下：

$$F_{f_adjust} = W_f \cdot (F_{c_adjust} + F_{t_adjust}) = W_f \cdot (W_c \cdot F_c + W_t \cdot F_t) \tag{6}$$

采用如上注意力机制对两种模态特征进行融合，可以自适应调整两种模态中的特征层权重，增强对决策有效特征的权重，弱化对决策无效特征的权重；并且该注意力机制可以随着网络训练进行同步优化，保证算法网络结构整体符合端到端的训练。

本文算法将设计好的特征融合模块密集地插入到多尺度特征层中，以将可见光和红外的特征层进行多尺度加权融合，一方面在多个特征层上进行特征融合；另一方面在特征融合时调整特征权值，提高多模态之间特征融合的有效性和充分性。

2 实验结果与分析

2.1 数据集与评价指标

本文采用 KAIST 数据集来验证所提算法的性能。KAIST 数据集中有 95 328 帧可见光与红外图像对，图片分辨率为 640×512，分为白天和夜间两种场景，其中包含光照条件恶劣、被不同程度遮挡、多尺度等复杂条件下的行人目标，数据集的标签分为 person、people 和 cyclist 三个类别。

本实验采用 8 963 对可见光与红外图像进行训练得出 CMMSNet 算法模型，用 2 252 对可见光与红外图像测试 CMMSNet 算法模型的性能。

采用平均漏检率(log-average miss rate)作为算法性能的评价指标，指标值越小表示模型性能越好。

2.2 实验环境及参数设置

本文算法采用的实验机器操作系统为 Ubuntu18.04，内存为 40 G，显卡型号为 NVIDIA Tesla V100，显存为 32 G。使用 Tensorflow+Keras 深度学习框架来实现本文算法。本文算法训练过程在 GPU 环境中进行，训练过程中算法主干特征提取网络采用 ResNet 网络结构，采用 Adam 优化器，模型训练的 epoch 次数为 300 次，batch_size 设置为 25，学习率初始值设置为 0.000 1，学习率衰减采用梯度下降的方式，即每隔 10 个 epoch 学习率缩小为原来的 1/10。

2.3 算法对比实验分析

本文算法 CMMSNet 与其他可见光红外融合行人检测算法对比结果如表 1 所示。通过表 1 可知，

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

ACF＋T＋THOG、Fusion RPN(region proposal network)、FFPN-Sharp、CIAN 等算法性能指标较差，因为它们对可见光和红外模态的特征只是在单一层上进行融合，很难有效融合两种模态的特征。针对以上算法在单一层上进行特征融合的缺陷，本文算法在多尺度的特征层上进行融合。实验结果表明本文算法可以有效、充分进行多模态特征融合，相比于对比算法，在行人检测任务上性能有较好的提升，算法具有良好的鲁棒性。

表 1　CMMSNet 算法与其他多模态行人检测算法的比较结果

算法	全天的平均漏检率/%	白天的平均漏检率/%	夜间的平均漏检率/%
ACF＋T＋THOG[9]	47.32	56.17	42.57
Fusion RPN[14]	35.28	36.77	33.91
FFPN-Sharp[11]	32.65	34.49	31.48
CIAN[13]	27.71	30.74	21.07
CMMSNet	19.98	22.95	14.32

同时，本文算法 CMMSNet 与单模态行人检测算法在 KAIST 数据集上的对比结果如表 2 所示。SSD-VIS 与 SSD-INF 分别是采用 SSD 目标检测算法在 KAIST 的可见光模态数据集和红外模态数据集上训练和测试的结果。通过表 2 可知，单独采用可见光和红外两种模态中的一种，训练出来的模型都很难满足复杂场景下的行人检测任务，所以需要将两种模态的特征进行融合来训练行人检测的模型。实验结果表明本文所提出的多模态行人检测算法相比单模态行人检测算法，在行人检测任务上性能有大幅度的提升。

表 2　CMMSNet 算法与单模态行人检测算法的比较结果

算法	全天的平均漏检率/%	白天的平均漏检率/%	夜间的平均漏检率/%
SSD-VIS(可见光)	59.53	50.32	77.62
SSD-INF(红外)	43.18	49.80	29.25
CMMSNet	19.98	22.95	14.32

以上结果表明，本文算法 CMMSNet 可以有效地融合可见光和红外模态图像特征。相比于单模态行人检测效果，CMMSNet 算法多模态行人检测的性能得到突破性的提升；并且相比于对比实验中的其他多模态行人检测算法，本文算法在复杂光照场景以及行人目标具有多尺度情况下能够实现更好的检测效果。

本文算法 CMMSNet 检测的效果，如图 5—图 6 所示。每张图片中左半部分为可见光图片，右半部分为红外图片，绿框为 CMMSNet 算法检测出的行人。可以看出 CMMSNet 算法在行人处于复杂光照场景以及行人具有多尺度情况下均有较好的鲁棒性。

图 5　CMMSNet 算法在白天的检测结果

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

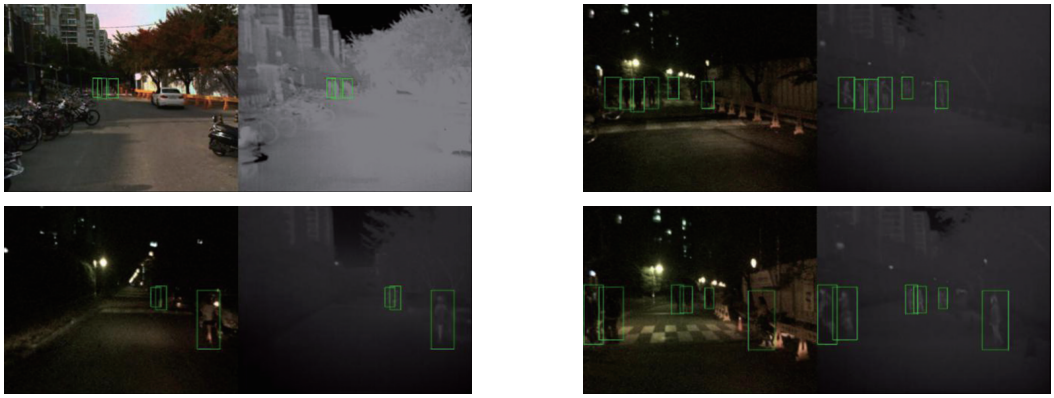

图 6 CMMSNet 在夜间或者光线阴暗条件下的检测结果

3 结 语

本文针对目前行人检测算法在复杂光照场景以及多尺度情况下检测效果较差的问题，提出了基于 SSD 多模态多尺度特征融合行人检测算法(CMMSNet)。相比于其他多模态行人检测算法只在单一特征层上进行特征融合，本文算法首先使用 ResNet 作为 SSD 算法中主干特征来提取网络，分别提取可见光图像和红外图像两种模态的特征，然后在多尺度的特征层上使用注意力机制对两种模态的特征进行加权融合，最后使用多尺度的特征融合层进行行人检测。在 KAIST 数据集上的实验结果表明，本文算法相较于单模态行人检测算法以及其他多模态行人检测算法在性能上有显著提升，在复杂光照场景以及多尺度情况下具有较好的鲁棒性。下一步计划研究算法模型的压缩，在不影响算法检测效果的情况下，通过降低算法模型的参数量，来提高算法的检测速度，增强算法的实用性。

参考文献：

[1] 蒋胜，黄敏，朱启兵，等. 基于 R-FCN 的行人检测方法研究[J]. 计算机工程与应用，2018，54(18)：180－183
[2] CAI Z，SABERIAN M，VASCONCELOS N. Learning complexity aware cascades for deep pedestrian detection[C]. Santiago：Proceedings of IEEE International Conference on Computer Vision，2015：3361－3369
[3] NGUYEN D T，LI W Q，OGUNBONA P O. Human detection from images and videos：a survey[J]. Pattern Recognition，2016，51：148－175
[4] 谢以磊，王军号，颜新云. 一种基于 MEAN-SHIFT 特征增强的车牌定位方法[J]. 金陵科技学院学报，2020，36(1)：16－20
[5] REN S，HE K，GIRSHICK R，et al. Faster R-CNN：towards realtime object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence，2015，39(6)：1137－1149
[6] LIU W，ANGUELOV D，ERHAN D，et al. SSD：single shot multibox detector[C]. Amsterdam：European Conference on Computer Vision，2016：21－37
[7] REDMON J，DIVVALA S，GIRSHICK R，et al. You only look once：unified，real-time object detection[C]. Boston：Proceedings of CVPR，2015：779－788
[8] ZHOU Z，DONG M，XIE X，et al. Fusion of infrared and visible images for night-vision context enhancement[J]. Appl Opt，2016，55(23)：6480－6490
[9] LIU Y，CHEN X，WANG Z，et al. Deep learning for pixel level image fusion：recent advances and future prospects[J]. Information Fusion，2018，42：158－173
[10] HWANG S，PARK J，KIM N，et al. Multispectral pedestrian detection：benchmark dataset and baseline[C]. Boston：Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2015：1037－1045
[11] 童靖然，毛力，孙俊. 特征金字塔融合的多模态行人检测算法[J]. 计算机工程与应用，2019，55(19)：214－222
[12] LIU J，ZHANG S，WANG S，et al. Multi-spectral deep neural networks for pedestrian detection[C]. Amsterdam：Proceedings of British Machine Vision Conference，2016：1－13
[13] ZHANG L，LIU Z，ZHANG S，et al. Cross-modality interactive attention network for multispectral pedestrian detection[J]. Information Fusion，2019，50：20－29
[14] KONIG D，ADAM M，JARVERS C，et al. Fully convolutional region proposal networks for multispectral person detection[C]. Hawaii：Proceedings of Computer Vision and Pattern Recognition Workshops，2017：243－250

(责任编辑：湛 江)

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net