引用格式：HAO Shuai, GAO Shan, MA Xu, et al. Infrared Pedestrian Detection Based on Cross-scale Feature Aggregation and Hierarchical Attention Mapping[J]. Acta Photonica Sinica, 2022, 51(6):0610006
郝帅,高山,马旭,等.基于跨尺度特征聚合与分层注意力映射的红外行人检测[J].光子学报,2022,51(6):0610006

# 基于跨尺度特征聚合与分层注意力映射的红外行人检测

郝帅[1],高山[1],马旭[1],安倍逸[1],何田[1],文虎[2],王峰[3]
(1 西安科技大学 电气与控制工程学院,西安 710054)
(2 西安科技大学 安全科学与工程学院,西安 710054)
(3 渭南师范学院 物理与电气工程学院,陕西 渭南 714000)

**摘　要**:针对红外行人图像中受多尺度、部分遮挡以及环境干扰导致传统算法难以准确检测问题,提出一种红外图像行人检测算法。以CSPdarknet53作为主干特征提取网络,在输入端引入Focus模块以减少主干网络下采样过程中小尺度目标特征信息丢失;通过构建跨尺度特征聚合模块来融合主干网络不同残差层输出的全局特征和多尺度局部特征,提高网络多尺度特征聚合能力,提升网络检测精度;针对红外图像受自身成像机理以及复杂背景影响造成行人目标特征难以有效表达的问题,通过构建分层注意力映射模块来增强行人特征表达能力。为了验证所提出算法的优势,选取4种经典对比算法,并在3种公共数据集上进行测试验证。实验结果表明,该算法可以实现复杂环境下多尺度红外行人的准确检测,其平均准确率和召回率分别可达95.37%和92.99%。

**关键词**:红外行人检测;多尺度;Focus模块;空间金字塔;注意力机制
**中图分类号**:TN215　　**文献标识码**:A　　**doi**:10.3788/gzxb20225106.0610006

## 0 引言

行人检测作为目标检测的一个重要分支,其主要任务是预测一系列包围行人的封闭边界框,从而标识出给定图像中的行人目标。目前,基于视觉的行人检测已经成为自动驾驶、智能监控以及机器人等领域的关键技术之一,同时也是计算机视觉领域的研究热点[1-2]。基于可见光图像的行人检测系统由于其易受光照变化等环境因素影响,尤其是在夜间或大雨大雾的情况下,很大程度上限制了其应用范围[3]。近年来,基于红外热成像的检测系统由于其抗干扰能力强、探测距离远、受光照变化影响小、能够全天候工作等优点,在行人检测、无人机侦查以及安防监控等领域得到了广泛应用。因此,开展基于红外热成像行人检测方法(红外行人检测方法)研究具有重要实际意义和理论研究价值[4-5]。

目前,红外行人检测方法主要分为两类:基于人工特征提取的机器学习方法和基于深度学习的检测方法。基于人工特征提取的机器学习方法主要通过构造人工特征和分类器判别相结合的方式实现,如先提取尺度不变特征变换(Scale-invariant Feature Transform, SIFT)[6]、方向梯度直方图(Histogram of Oriented Gradient, HOG)[7]、哈尔(Haar)特征[8]等特征,再利用支持向量机(Support Vector Machines, SVM)[9]或Adaboost[10]等分类器进行分类识别从而得到检测结果。BISWAS S K等[11]提出了一种基于SVM的热红外行人检测算法,并通过实验证明了该方法可以在噪声条件下实现行人目标的准确检测。然而,当行人目标

**基金项目**:国家自然科学基金(No. 51804250),中国博士后科学基金(Nos. 2019M653874XB, 2020M683522),陕西省科技计划项目(Nos. 2021JQ-572, 2020JQ-757),陕西省教育厅科研计划项目(No. 18JK0512),渭南市科技计划项目(No. 2020ZDYF-JCYJ-196),陕西省创新能力支撑计划(No. 2020TD-021),西安市碑林区科技计划项目(No. GX2116)
**第一作者**:郝帅(1986-),男,讲师,博士,主要研究方向为人工智能、模式识别。Email:haoxust@163.com
**通讯作者**:马旭(1985-),女,讲师,博士,主要研究方向为人工智能及机器视觉。Email:414548542@qq.com
**收稿日期**:2021-12-13;**录用日期**:2022-02-22

较为密集或存在部分遮挡时，该方法的漏检以及误检情况较为严重。HIRANMAI M等[12]利用加速鲁棒性特征（Speeded Up Robust Feature，SURF）结合SVM分类器实现行人检测，通过实验证明了该方法可以提高部分遮挡条件下行人目标的检测性能，但对于边缘模糊、纹理特征较弱的红外图像，检测精度较低。LIU Yande等[13]提出了一种结合Oriented-fast and Rotated Brief（ORB）特征和HOG特征的级联两阶段分类行人检测方法，并通过实验证明该方法可以提升复杂环境下的行人检测精度，但实时性欠缺。综上，基于机器学习的方法虽然可以满足某些特定环境下的检测要求，但由于人工提取特征时依赖于专家经验，且手工特征抗干扰能力弱，往往存在实时性不强、泛化能力弱、鲁棒性差等问题。

近年来，基于深度学习的目标检测算法相比于传统机器学习算法在检测精度和实时性方面展现出明显优势[14-17]。目前，基于深度学习的目标检测算法主要分为two-stage和one-stage两类。two-stage算法的主要思想是先寻找候选区域，然后在候选区域上对检测结果分别进行分类和位置回归，具有代表性的算法包括基于候选区域的区域卷积神经网络（Region-based Convolutional Neural Network，R-CNN）[18]、Fast R-CNN[19]以及Faster R-CNN[20]等。LI Jianan等[21]对Faster R-CNN算法进行改进，通过构建两种子网络分别用以检测大尺度和小尺度行人目标并将预测结果加权融合来提高检测模型的多尺度检测性能，但该方法没有充分利用到多尺度特征图的局部区域特征，因此对遮挡区域的检测效果依然不足。ZHANG Liliang等[22]在Faster R-CNN基础上采用孔算法（hole algorithm）增加特征图尺度，以提升小尺度目标检测效果，同时采用级联增强森林算法对正负样本重新加权，以减少复杂背景对目标检测的干扰。然而，two-stage方法在确定候选区域时需要一定时间，实时性普遍较差，且候选框之间存在大量重叠，提取特征操作冗余，占用大量存储空间。

One-stage目标检测模型代表算法包括Single Shot Multibox Detector（SSD）[23]和You Only Look Once（YOLO）系列[24-27]。此类算法是将目标检测过程看作一个回归问题，以整张图片作为网络输入，直接在输出层对边界框的位置和类别进行回归，相较于第一类算法很大程度提高计算速度。赵斌等[28]采用改进的YOLOv3算法进行红外行人检测，并通过实验证明了该方法在检测精度和速度上均优于Faster RCNN和SSD两种算法。但是该方法对于多目标相互遮挡、重叠区域的处理能力较弱。WEN B Y等[29]利用YOLOv4算法在Caltech数据集上进行行人检测实验，并通过实验验证了该方法相比于YOLOv3算法具有更高的检测精度。DU S J等[30]利用YOLOv4实现了复杂环境下的红外图像车辆检测，通过实验证明其相比于其他检测网络具有更好的实时性。

基于YOLOv4的目标检测算法可以较好地兼顾目标检测精度和速度。然而，当利用YOLOv4进行红外图像的行人检测时，仍然存在问题：1）基于YOLOv4的主干特征提取网络由于存在较多的下采样运算易导致底层特征描述能力不足，造成小尺度目标的检测性能受限；2）在复杂场景中，多目标之间相互遮挡以及行人目标所存在的多尺度特点也会对YOLOv4检测的性能提出挑战；3）由于红外图像自身具有纹理特征弱、空间分辨力差等缺点，行人目标易被高亮背景所淹没，进而导致YOLOv4网络难以准确定位目标区域。

针对上述问题，本文在YOLOv4目标检测网络基础上提出一种基于跨尺度特征聚合与分层注意力映射（Cross-scale Feature Aggregation and Hierarchical Attention Mapping，CFAHAM）的多尺度红外行人检测方法。在主干特征提取网络中引入Focus模块，利用切片分割的采样方式对输入图像进行无损降采样，从而减少主干网络下采样过程中的信息丢失，提高小尺度目标检测性能。设计了一个跨尺度特征聚合模块，通过构建多尺度空间金字塔池化层对主干输出特征进行分区池化，使网络在训练过程中可以融合不同尺度的行人特征，从而改善多尺度以及部分遮挡情况下的行人检测效果。构建分层注意力映射模块，使检测器在特征提取的过程中快速聚焦于行人目标，从而有效增强复杂环境下行人检测性能。

## 1 CFAHAM红外行人检测算法

CFAHAM红外行人检测算法流程如图1所示。

搭建的检测算法框架流程为：1）将数据集划分为训练集和测试集；2）对训练集行人样本进行标注并将标签数据传入训练网络；3）设置训练参数，首先将图像大小调整为416×416进行第一阶段预训练得到预训练权重，然后再将图像尺寸调整为608×608进行第二阶段再训练；4）训练过程中，为了提高小尺度目标检测

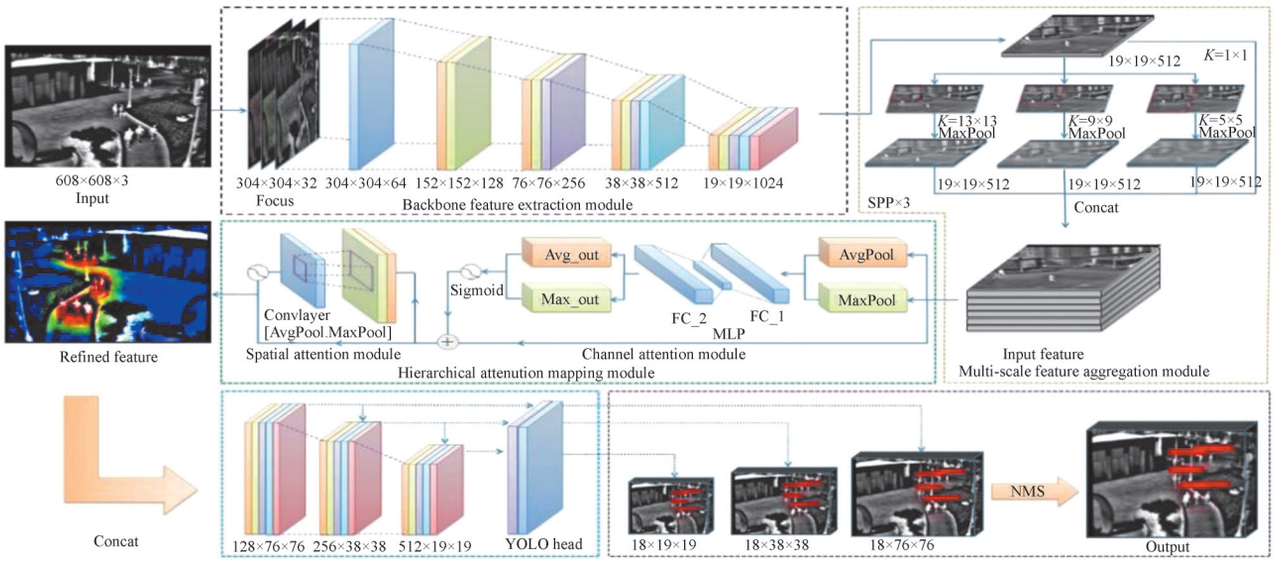



图1　检测算法流程

Fig. 1　Flow chart of detection algorithm

性能，在输入端CSPdarknet53主干特征提取网络上引入Focus切片采样模块，将输入图像的空间维度信息提取到通道维度，从而得到更加完备的下采样特征信息，为后续小尺度目标特征的有效提取奠定基础；5）为了改善因遮挡不同造成的目标特征形变和视野深度不同造成的尺度多样性问题，在主干特征提取网络与特征传递网络之间构建基于空间金字塔池化的跨尺度特征聚合模块，以提高模型对特征形变和多尺度目标的检测鲁棒性；6）为了提高复杂环境下行人检测精度，在特征金字塔的多个特征传递分支上引入视觉注意力机制，设计出一种分层注意力映射模块，用以增强复杂环境下行人特征显著性的同时抑制背景信息，增强检测模型对人体特征的感知能力；7）训练完成后，利用得到的训练模型权重在测试集中进行测试验证。本文算法训练过程伪代码如下：

算法：CFAHAM红外行人检测算法

输入：主干网络输入图像：$\boldsymbol{X}_{\text{input}}$，训练周期 $T_{\text{epoch}}$，训练批次大小 $\boldsymbol{S}_{\text{batch}}$，初始学习率 $\boldsymbol{\eta}_{\text{initial}}$

While（$T<T_{\text{epoch}}$）do

1. 主干特征提取网络

1）利用Focus+CBM结构进行降采样和维度扩展

$$\boldsymbol{X}_{\text{extend}}=\text{CBM}(\text{Focus}(\boldsymbol{X}_{\text{input}}))$$

2）通过CSPdarknet53网络的Resblock残差结构进一步提取主干特征信息

$$\boldsymbol{X}_{\text{out\_1}}[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3]=\text{Resblock}(\boldsymbol{X}_{\text{extend}})$$

2. 跨尺度特征聚合模块

输入：主干网络输出特征图 $\boldsymbol{X}_{\text{out\_1}}[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3]$；初始化：3尺度最大池化网络 $F_{\max}(f_1,f_2,f_3)$

for：$f_\alpha,\boldsymbol{x}_\beta$ in $F_{\max}(f_1,f_2,f_3),\boldsymbol{X}_{\text{out\_1}}[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3]$ do

1）通过多尺度池化核计算分区池化特征图：

$$[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3]=f_\alpha(\boldsymbol{x}_\beta)$$

2）将输出特征图在通道维度进行拼接，并使用多层卷积结构进行通道整合，实现数据降维：

$$\boldsymbol{X}_{\text{out\_2}}[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3]=\text{Conv}\{\text{Concate}[[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3]+\boldsymbol{x}_\beta,\text{dim}=1]\}$$

end for

3. 分层注意力映射模块

输入：跨尺度特征聚合模块输出特征图 $\boldsymbol{X}_{\text{out\_2}}[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3]$；初始化：分层注意力映射模块HAM(•)

$$F_1=\mathrm{HAM}\left(\left[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3\right],\ \mathrm{size}=\boldsymbol{X}_{1.\mathrm{size}}\right)$$

$$F_2=\mathrm{HAM}\left(\left[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3\right],\ \mathrm{size}=\boldsymbol{X}_{2.\mathrm{size}}\right)$$

$$F_3=\mathrm{HAM}\left(\left[\boldsymbol{X}_1,\boldsymbol{X}_2,\boldsymbol{X}_3\right],\ \mathrm{size}=\boldsymbol{X}_{3.\mathrm{size}}\right)$$

4. 预测层

1)CIOU损失函数：

$$L_{\mathrm{CIOU}}=1-R_{\mathrm{IU}}+\frac{\rho^2\left(b,b_{\mathrm{gt}}\right)}{c^2}+\frac{v^2}{\left(1-R_{\mathrm{IU}}\right)+v}$$

$$R_{\mathrm{IU}}=\frac{S\cap S_{\mathrm{gt}}}{S\cup S_{\mathrm{gt}}},v=\frac{4}{\pi^2}\left(\arctan\frac{w_{\mathrm{gt}}}{h_{\mathrm{gt}}}-\arctan\frac{w}{h}\right)^2$$

2)反向传播梯度：loss.backward()

3)学习率更新：

$$\eta_{\mathrm{t}}=\eta_{\min}+\frac{1}{2}\left(\eta_{\mathrm{initial}}-\eta_{\min}\right)\left[1+\cos\left(\frac{\mathrm{epoch}}{T}\pi\right)\right]$$

End

输出：最优训练权重

### 1.1 YOLOv4目标检测模型

YOLOv4目标检测算法在YOLOv3算法框架下，通过在输入端、主干网络、损失函数等方面进行改进，能够在具有较高检测精度的同时兼顾算法的实时性。

输入端：首先将不同尺寸的输入样本图像统一调整为416×416，在缩放填充的过程中对原始样本自适应填充，然后再送入检测网络的Backbone中。另外，通过对数据集采用马赛克(Mosaic)数据增强、自对抗训练(Self-adversarial-Training，SAT)等方法丰富训练样本，提高网络的鲁棒性和泛化迁移能力。

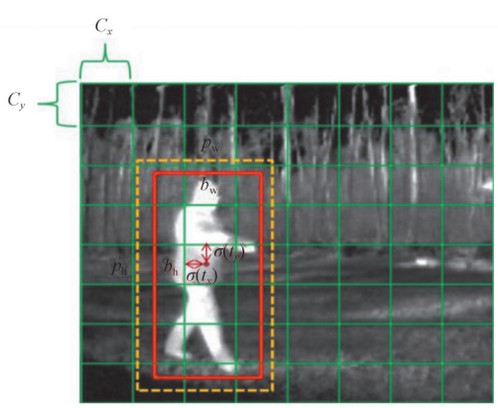
主干特征提取网络：为了使检测模型轻量化的同时兼顾检测精度，YOLOv4主干特征提取网络采用了CSPDarknet53网络结构，如图2(a)所示。

颈部网络：为了融合多尺度目标特征，YOLOv4算法在传统的特征金子塔网络(Feature Pyramid Networks，FPN)基础上使用了自底向上的金字塔注意力网络(Pyramid Attention Network，PAN)结构，两者结合可以增强检测器特征聚合能力，如图2(c)所示。

预测端：在预测时YOLOv4目标检测算法首先通过FPN-PAN特征金字塔融合网络中不同尺度的特征图，然后在此基础上利用三尺度的输出特征图进行边界框的预测，最后将重合度高于阈值的边界框利用非极大值抑制算法滤除，从而得到最终检测结果，预测过程如图2(b)所示。

| | Type | Filters | Size | Output |
|---|---|---|---|---|
| | Convolutional | 32 | 3×3 | 608×608 |
| 1× | Downsample_conv | 64 | 3×3 | 304×304 |
| | Split-conv | 64 | 1×1 | 304×304 |
| | Resblock | | | 304×304 |
| | Concat_conv | 64 | 1×1 | 304×304 |
| 2× | Downsample_conv | 128 | 3×3 | 152×152 |
| | Split-conv | 64 | 1×1 | 152×152 |
| | Resblock | | | 152×152 |
| | Concat_conv | 128 | 1×1 | 152×152 |
| 8× | Downsample_conv | 256 | 3×3 | 76×76 |
| | Split-conv | 128 | 1×1 | 76×76 |
| | Resblock | | | 76×76 |
| | Concat_conv | 256 | 1×1 | 76×76 |
| 8× | Downsample_conv | 512 | 3×3 | 38×38 |
| | Split-conv | 256 | 1×1 | 38×38 |
| | Resblock | | | 38×38 |
| | Concat_conv | 512 | 1×1 | 38×38 |
| 4× | Downsample_conv | 1024 | 3×3 | 19×19 |
| | Split-conv | 512 | 1×1 | 19×19 |
| | Resblock | | | 19×19 |
| | Concat_conv | 1024 | 1×1 | 19×19 |

(a) CSPDarknet53 network structure




(b) Bounding box prediction

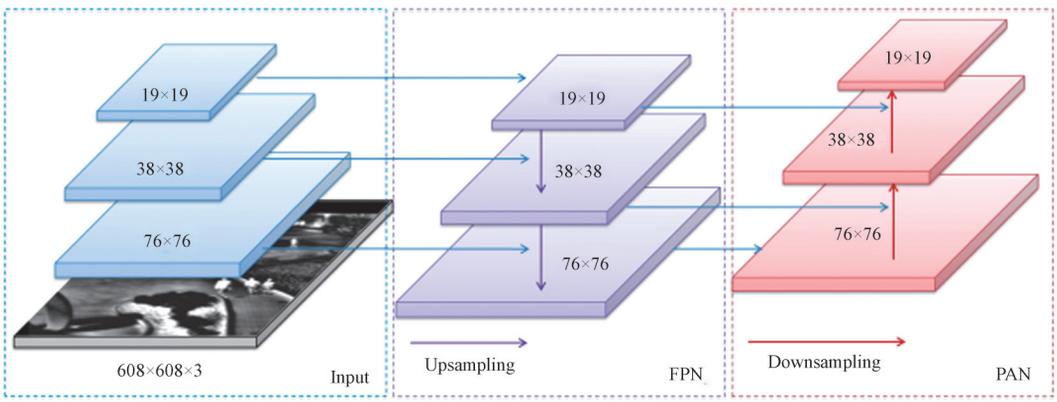



图 2　YOLOv4 目标检测算法

Fig. 2　YOLOv4 object detection algorithm

图 2(b)中各参数定义为

$$b_x = \sigma(t_x) + c_x \tag{1}$$

$$b_y = \sigma(t_y) + c_y \tag{2}$$

$$b_w = p_w \exp(t_w) \tag{3}$$

$$b_h = p_h \exp(t_h) \tag{4}$$

式中，$b_x$ 和 $b_y$ 分别为边界框中心点横、纵坐标；$b_w$ 和 $b_h$ 分别为边界框的宽和高；$\sigma(t_x)$ 和 $\sigma(t_y)$ 分别表示预测框中心点偏离其所在网格左上角水平方向和垂直方向的距离，$\sigma(\cdot)$ 表示 sigmoid 函数；$c_x$ 和 $c_y$ 分别为每个网格与图像左上角的横纵坐标距离；$p_w$ 和 $p_h$ 分别为先验框的宽度和高度；$t_w$ 和 $t_h$ 分别为边界框的横、纵方向尺度缩放因子。

在训练过程中，采用余弦退火衰减算法使学习率随损失收敛以余弦函数的方式周期性调整，数学表达式为

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\text{initial}} - \eta_{\min})\left[1 + \cos\left(\frac{\text{epoch}}{T}\pi\right)\right] \tag{5}$$

式中，$\eta_t$ 表示新的学习率；$\eta_{\text{initial}}$ 表示初始学习率；$\eta_{\min}$ 表示最小学习率，取为 0.000 1；epoch 表示训练迭代次数；$T$ 表示余弦函数的半周期，取为 5。

为了提高训练过程中预测框的回归速度和精度，同时避免损失函数在训练过程中出现梯度发散问题，采用完全交并比(Complete Intersection Over Union，CIOU)作为模型训练损失函数 $L_{\text{CIOU}}$，定义为

$$L_{\text{CIOU}} = 1 - R_{\text{IU}} + \frac{\rho^2(b, b_{\text{gt}})}{c^2} + \frac{v^2}{(1 - R_{\text{IU}}) + v} \tag{6}$$

$$R_{\text{IU}} = \frac{S \cap S_{\text{gt}}}{S \cup S_{\text{gt}}}, v = \frac{4}{\pi^2}\left(\arctan\frac{w_{\text{gt}}}{h_{\text{gt}}} - \arctan\frac{w}{h}\right)^2 \tag{7}$$

式中，$R_{\text{IU}}$ 表示预测框与真实框面积的交并比；$\rho(\cdot)$ 表示欧氏距离；$b$ 和 $b_{\text{gt}}$ 分别表示预测框与真实框中心点；$c$ 表示包含预测框与真实框最小外接矩形的对角线距离；$v$ 表示衡量预测框和真实框宽高比一致性的参数；$S$ 和 $S_{\text{gt}}$ 分别表示预测框和真实框的面积；$w_{\text{gt}}$、$h_{\text{gt}}$ 和 $w$、$h$ 分别表示真实框和预测框的宽和高。

### 1.2　Focus 切片采样原理

由于图像中行人目标位置不同从而使得行人目标在图像中的视野深度也不同。红外图像数据集中存在大量分布不均的小尺度行人目标，而主干网络的下采样易导致小尺度目标信息丢失从而影响检测器性能。Focus 模块可以对图像进行切片采样，通过对图像的每个通道像素值进行间隔采样，将原始图像中的空间维度信息提取到通道维度，从而得到更加完备的下采样目标特征信息，如图 3 所示。

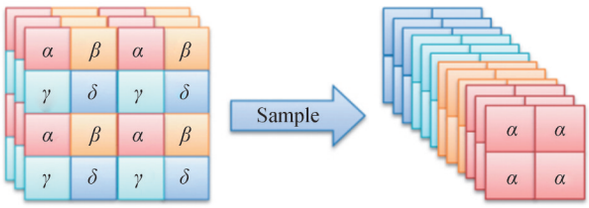



图3 Focus切片采样原理

Fig.3 Focus slice sampling principle

此外，由于YOLOv4算法在图像输入检测网络后采用CSPResblock_1残差结构进行下采样和维度扩充，易造成小尺度行人特征信息丢失。为此，对检测网络的输入端进行改进，将原始的CSPResblock_1结构替换为Focus+CBM(Conv2D + BatchNormalization + Mish激活函数)结构，先利用切片分割的采样方式减少主干网络下采样过程中带来的信息丢失，提高小尺度目标的检测性能；然后通过CBM模块进行通道维度整合，用以连接CSPResblock_1残差层输入，以便实现小尺度目标信息的充分表达，网络结构如图4所示。

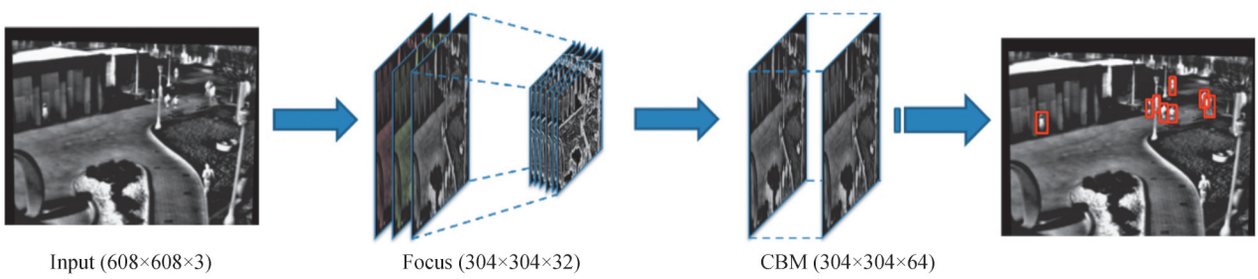



图4 Focus+CBM网络结构

Fig.4 Focus+CBM network structure

### 1.3 基于空间金字塔池化的跨尺度特征聚合模块

YOLOv4算法通过使用非极大值抑制算法来解决同类目标被多次检测的问题，但由于行人图像中可能存在因目标遮挡程度不同导致行人特征形变甚至丢失的问题，仅依靠非极大值抑制算法容易导致漏检和误检现象。人类视觉通路中由于存在同层多尺度感受野，可以使人类视觉系统同时感知不同尺度的物体，从而对尺度变化进行自适应调整，以达到最佳的认知效果。基于上述思想，通过模拟大脑视觉通路的同皮层多尺度感受野认知机理，在主干特征提取网络与特征传递网络之间引入基于空间金字塔池化的跨尺度特征聚合模块，从而扩展网络主干特征感受野并提高网络多尺度特征融合能力，进而改善检测模型的多尺度和部分遮挡区域的检测性能。

空间金字塔池化用来融合多尺度的局部区域特征，如图5所示。首先将不同尺度的输入特征图划分为

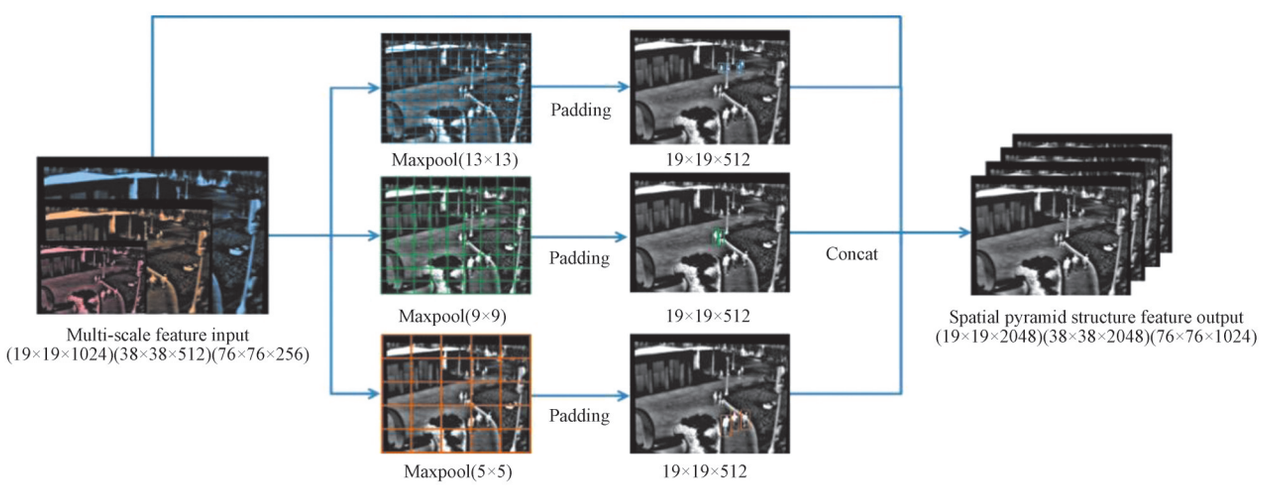



图5 空间金字塔池化结构

Fig.5 Spatial pyramid pooling structure

$a_i = n_i \cdot n_i$个数的空间窗口（$a_i$表示第$i$层特征图的窗口个数；$n_i$表示最大池化核大小，分别取5、9、13），然后通过3尺度池化核的最大池化计算分区池化特征图，扩展网络主干特征感受野；最后，对输出的多尺度池化特征图进行填充，使其具有空间尺度一致性，并在通道维度进行拼接，从而实现多尺度的局部区域特征聚合。

将跨尺度特征聚合模块构建在主干特征提取网络和多层特征传递网络之间，如图6虚线部分所示。首先，在主干特征提取网络输出的3尺度特征图后构建多层空间金字塔池化网络，有助于利用网络不同残差层输出的全局特征融合多尺度特征实现跨尺度特征聚合，进而提高模型的多尺度检测性能；其次，在3尺度的残差层之后利用空间金字塔网络对每个残差层输出特征图进行分区最大池化和特征聚合，从而结合单个特征图全局空间特征信息和多尺度局部区域特征信息，进而提升模型对于多尺度空间布局和物体形变的鲁棒性，以解决因行人遮挡程度不同造成的特征形变和视野深度不同造成的尺度多样性问题；最后，考虑到CSPResblock_8残差层输出的特征图具有最高空间分辨率，而高空间分辨率有利于小尺度目标检测，因此在该尺度构建金字塔池化层进行局部区域特征融合有助于弥补算法对小尺度目标检测性能不足的缺陷。

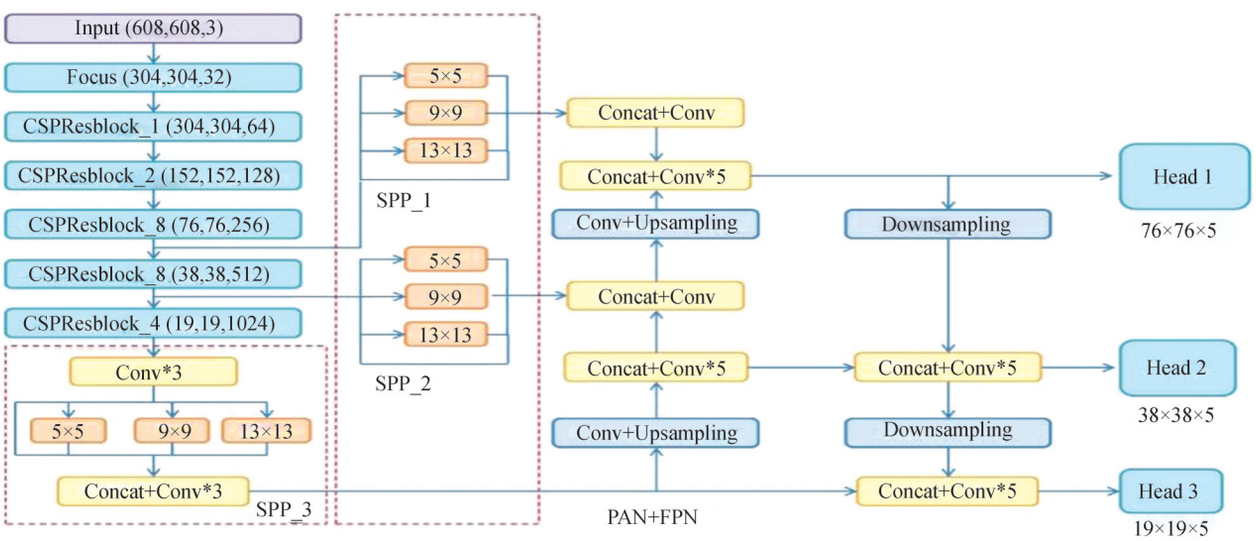



图6　基于空间金字塔池化的跨尺度特征聚合模块

Fig.6　Multi-scale feature aggregation module based on spatial pyramid pooling

### 1.4　基于CBAM的分层注意力映射模块

针对红外图像受复杂背景干扰导致多尺度行人目标显著度较弱进而造成传统算法检测精度低的问题，受视觉神经分层认知机制的启发，在特征金字塔融合网络中引入了分层注意力映射模块，通过在FPN-PAN特征金子塔的多个特征传递分支结构上嵌入卷积块注意力模型（Convolutional Block Attention Module，CBAM）[31]，使网络在多层特征传递分支上进行特征筛选，从而捕捉不同层次多尺度行人特征之间的相关性，同时从空间和通道两个维度提高复杂环境下行人特征的显著性，增强检测模型对行人特征的感知能力，其网络结构如图7所示。

图7中，将CSPdarknet53主干输出特征按照由浅到深划分为3层，底层维度（76，76，256），中层维度（38，38，512），高层维度（19，19，1024）。网络底层具有高的空间分辨率，建立行人目标定位特征的注意力机制，从而确定基本感兴趣区域；在中层网络构建行人目标表观特征注意力机制，如目标图像对比度、行人轮廓、尺度等；网络的高层输出特征具有最广阔的视觉感受野，用来建立行人目标的语义特征注意力机制。通过3层具有不同目的导向的特征筛选，从而可以更精细地确定行人目标的特征信息。

CBAM模块在结构上依次结合通道注意力模块和空间注意力模块。通过在不同特征层上依次沿通道和空间两个独立维度进行注意力映射，并分别采用全局平均池化（Average Pooling）和最大池化（Max Pooling）对目标区域的特征信息进行增强，通道注意力机制对不同尺度特征图之间的相关性进行建模，为每个通道赋予不同的权重系数强化行人的语义特征。计算公式为

$$\boldsymbol{M}_c(\boldsymbol{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\boldsymbol{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\boldsymbol{F}))) = \sigma(\boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{S}})) + \boldsymbol{W}_1(\boldsymbol{W}_0(\boldsymbol{F}_{\mathrm{max}}^{\mathrm{S}}))) \quad (8)$$

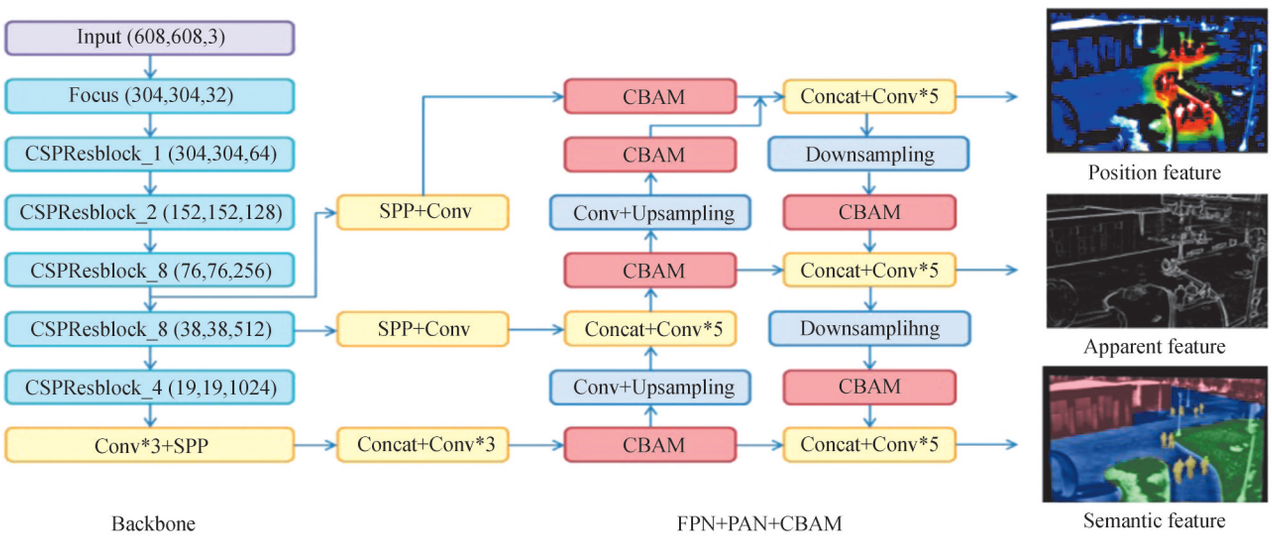



图7 基于CBAM的分层注意力映射模块

Fig.7 Hierarchical attention mapping module based on CBAM

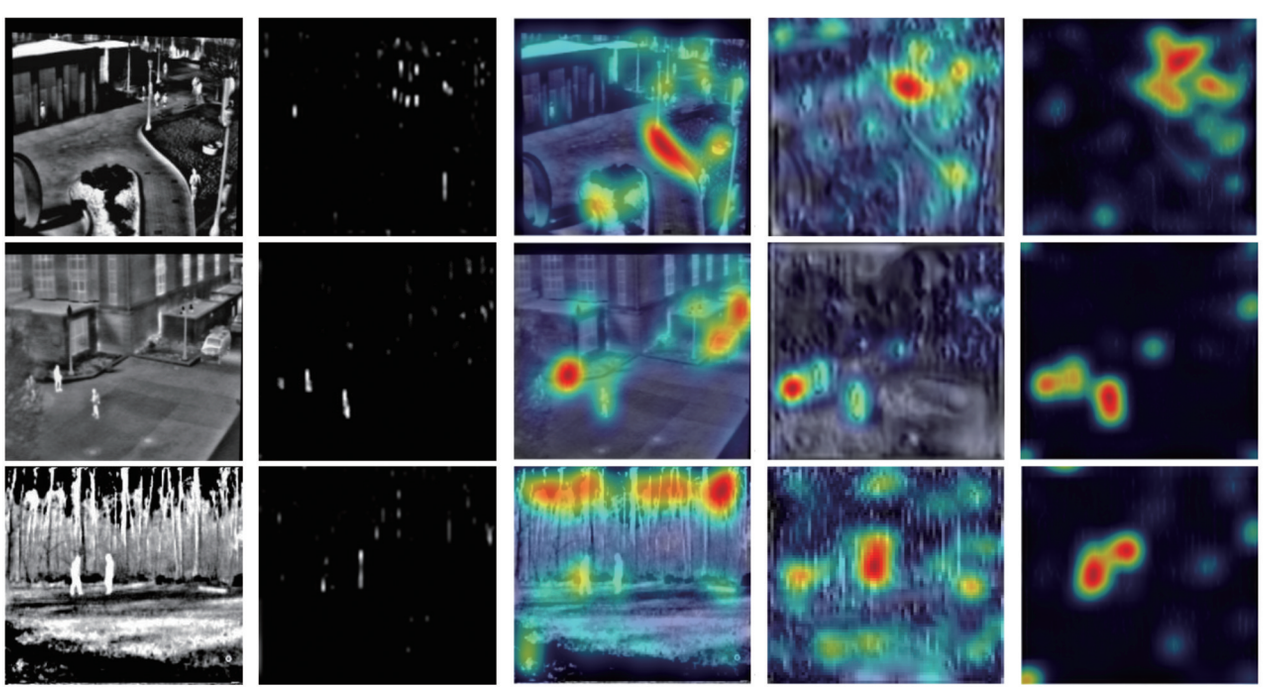
式中，$\boldsymbol{M}_{\mathrm{c}}(\boldsymbol{F})$表示通道注意力机制输出特征图；$\sigma(\cdot)$为sigmoid非线性激活函数；MLP表示CBAM中的共享网络；AvgPool(·)和MaxPool(·)分别表示平均池化和最大池化；$\boldsymbol{W}_0$和$\boldsymbol{W}_1$分别表示MLP中多层感知器中的隐藏层权重和输出层权重；$\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{S}}$和$\boldsymbol{F}_{\mathrm{max}}^{\mathrm{S}}$分别表示空间全局平均池化特征和最大池化特征。

空间注意力机制利用特征间的空间关系，对位置信息生成权重掩膜并加权输出，生成空间显著性特征图，提升关键区域特征表达，弱化背景等不相关区域，增强行人目标定位特征，计算过程为

$$\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F})=\sigma(f_{7\times7}([\mathrm{AvgPool}(\boldsymbol{F}));\ \mathrm{MaxPool}(\boldsymbol{F})]))=\sigma(f_{7\times7}([\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{C}};\ \boldsymbol{F}_{\mathrm{max}}^{\mathrm{C}}]) \tag{9}$$

式中，$\boldsymbol{M}_{\mathrm{s}}(\boldsymbol{F})$表示空间注意力机制输出特征图；$f_{7\times7}$表示7×7的卷积运算，$\boldsymbol{F}_{\mathrm{avg}}^{\mathrm{C}}$和$\boldsymbol{F}_{\mathrm{max}}^{\mathrm{C}}$分别表示通道全局平均池化特征和最大池化特征。

图8为输入图像显著性分布和注意力映射模块输出图像显著性分布对比。图8(a)为输入图像，选取了4种典型场景进行测试，从上到下依次为多尺度和小尺度目标场景、遮挡场景、复杂背景和高亮伪目标场景。图8(b)为显著系数图，表征特征图不同区域显著性，在图中行人及部分高亮物体具有较高显著性，呈高亮矩



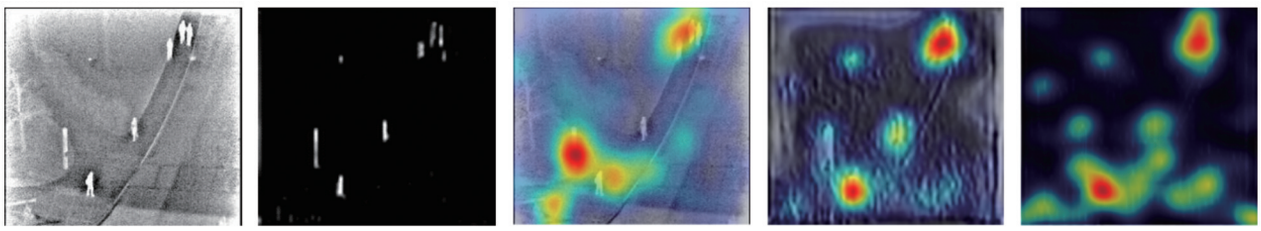



图8　显著性分布对比

Fig.8　Comparison of saliency distribution

形，与目标无关的背景区域在图中表现为黑色区域，说明其特征响应被有效抑制。图8(c)为输入图像显著性分布图，图8(d)、(e)分别为注意力映射模块输出显著性分布图和其边缘显著性分布图。

对比图8(c)、(d)显著性分布图可以看出，在原始输入图像中，道沿、屋檐、树干等边缘区域以及车辆、路灯等高亮物体显著性较高，而在注意力映射模块输出图中这些区域的特征响应被显著降低，同时，行人目标区域显著性被逐渐增强。从而说明本文所设计的注意力映射模块可以有效抑制输入图像中的复杂背景因素干扰，增强行人特征的显著性。此外，从图8(e)可以看出，注意力映射模块使检测模型聚焦于特定的局部区域，为后续行人目标的精确检测奠定基础。

## 2　实验与数据分析

实验环境配置参数如表1所示。

**表1　实验环境配置参数**

**Table 1　Experimental environment configuration parameters**

| Configuration | Version parameters |
|---|---|
| GPU | Nvidia GeForce GTX 1080Ti |
| CPU | Intel(R) Core(TM) i7-8700K 3.70GHz @2.90GHz×6 CPUs |
| Operating system | Microsoft Windows 10 |
| Deep learning framework | Pytorch 1.2.0 CUDA 10.0 |

### 2.1　实验数据集

实验选用OTCBVS公共基准数据库中的OSU Color-Thermal Database[32]、Terravic Motion IR Database和OSU Thermal Pedestrian Database[33]3个红外行人检测数据集。利用OSU Color-Thermal Database数据集对模型进行训练，该数据集包含17 089张红外和可见光视频图像序列，图像大小为320×240。为了验证本文算法的泛化能力，利用其它两个数据集进行测试。

### 2.2　数据预处理和模型训练

在输入端，为了丰富训练样本，采用Mosaic数据增强的方式对原始数据进行裁剪、缩放、拼接，从而实现数据集的扩充。同时，为了使模型适应不同尺度的输入图像，提高小尺度目标的检测精度，将训练过程分为两个阶段。第一阶段将输入图像大小调整为416×416进行100个批次的训练(学习率取为0.01)，从而得到红外数据集上的预训练权重。第二阶段再将输入图像大小调整为608×608进行再训练(学习率取为0.001)。在训练初期(0～50 epoch)对主干网络进行冻结可以防止预训练权值被修改并加快训练速度，第二阶段对整个网络的参数进行训练。

### 2.3　实验结果与对比分析

在YOLOv4算法基础上通过引入Focus模块，并设计了跨尺度特征聚合模块和分层注意力映射模块来提升检测模型对于红外图像中的行人检测性能。因此，为了验证各模块作用，进行了消融实验，其结果如表2所示。

如表2第1行所示，YOLOv4目标检测算法在测试集上的平均准确率为87.46%。分析不同改进措施的

表2 不同策略下的检测结果对比

Table 2 Comparison of detection results of different strategies

| K-means | SPP | CBAM | Focus | AP/% | Precision/% | Recall/% | F1 |
|---|---|---|---|---|---|---|---|
| √ | | | | 87.46 | 91.79 | 77.09 | 0.84 |
| √ | √ | | | 91.48 | 87.43 | 85.87 | 0.87 |
| √ | | √ | | 94.77 | 92.72 | 87.88 | 0.90 |
| √ | | | √ | 94.21 | 89.69 | 91.88 | 0.91 |
| √ | √ | √ | √ | **95.37** | **94.25** | **92.99** | **0.94** |

性能评估结果可以得出，几种改进网络的平均检测准确率相较于YOLOv4算法均有不同程度的提升，且本文算法通过融合多种改进措施后，检测精度相比原YOLOv4算法提升7.91%。

为了验证本文算法对于多尺度目标的检测效果，在OSU数据集中进行了测试，其检测结果尺度分布对比如图9所示。图中横、纵坐标分别为行人目标宽、高尺度，蓝色标记为YOLOv4算法检测结果尺度分布，红色标记为改进算法检测结果尺度分布。可看出，YOLOv4算法对于图表左下角小尺度行人目标存在较多漏检，而本文算法在小尺度目标区域具有更优秀的检测效果，证明采用Focus+CBM结构可以有效提升检测模型的小尺度目标检测性能。

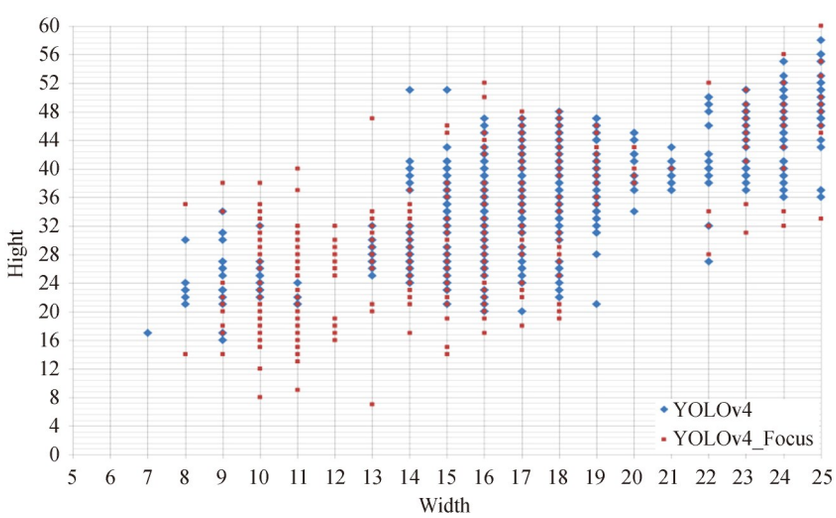
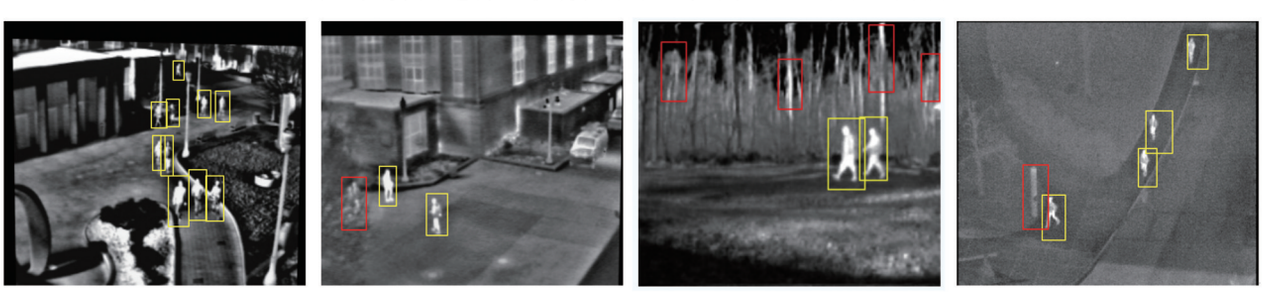


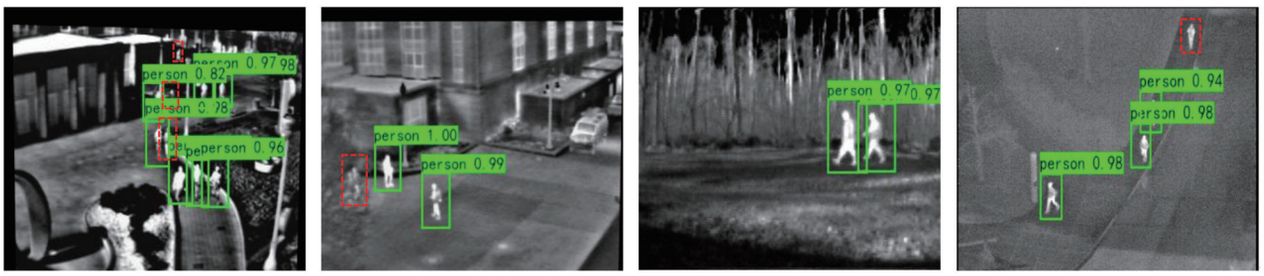

图9 检测结果尺度分布对比

Fig.9 Comparison of scale distribution in test results

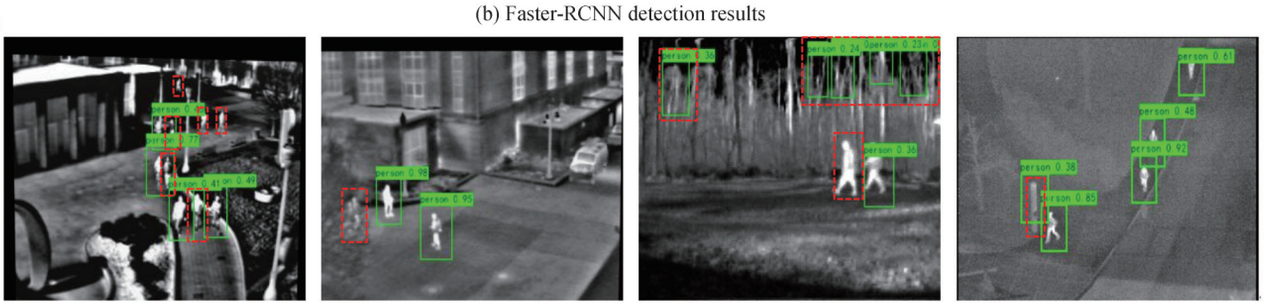
为了验证本文算法在各种复杂场景下的检测性能，选取了4种典型场景进行测试验证，如图10(a)所示，从左到右分别为存在多尺度目标情况、存在遮挡情况、存在多个高亮伪目标以及存在亮度及形状相似的情况，黄色框表示真实行人目标，红色框表示容易出现漏检及误检的区域。实验中选取了Faster RCNN、SSD、YOLOv3以及YOLOv4 4种经典算法进行比较，算法对比结果如图10(b)～(f)所示。



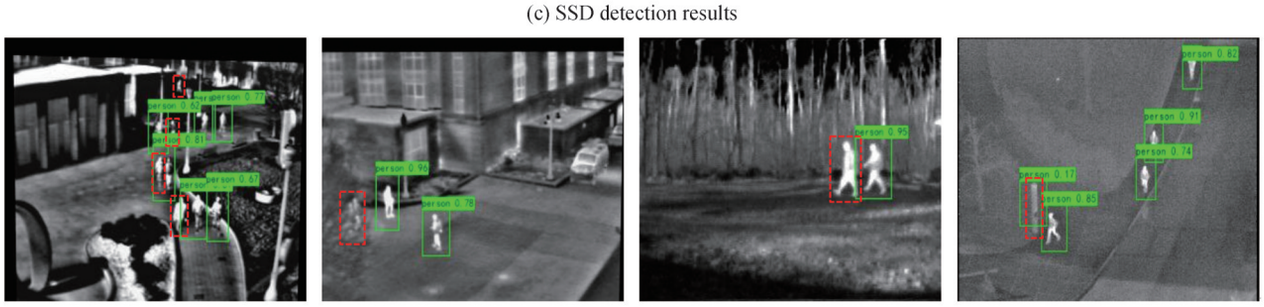
(a) Original image



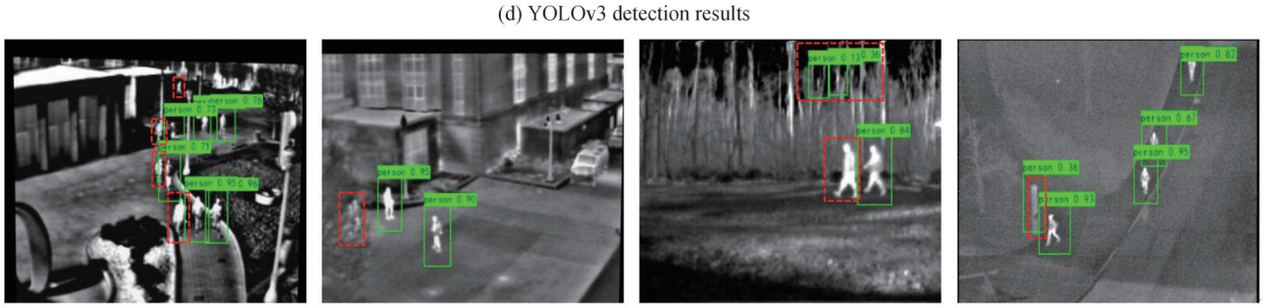

(b) Faster-RCNN detection results



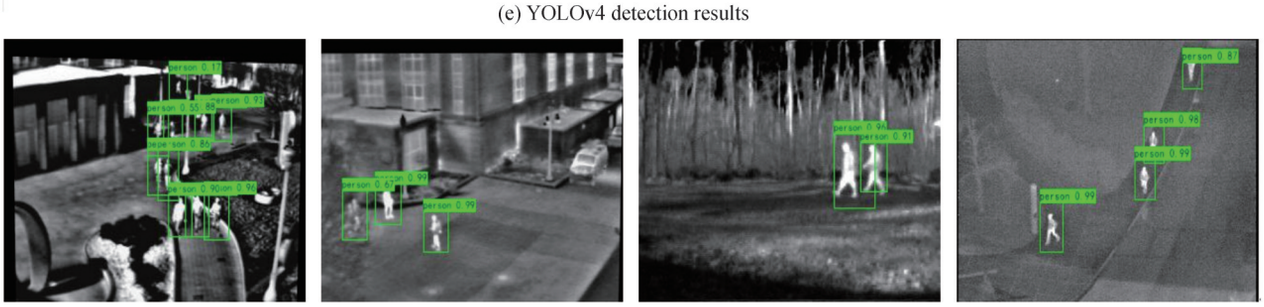
(c) SSD detection results



(d) YOLOv3 detection results



(e) YOLOv4 detection results



(f) CFAHAM detection results

图 10　可视化检测结果对比

Fig. 10　Comparison of visualization detection results

图 10(b)～(f)中的绿色框为算法预测框，对于漏检和误检区域采用红色虚线框进行标注。从图 10 第 1 列图像可以看出，4 种对比算法均存在漏检，无法检测出小尺度目标。此外，图像中间区域中行人目标存在部分遮挡，各对比算法均未检测出。而本文算法引入 Focus 模块减少了小尺度目标特征信息丢失，从而可以较好地检测出小尺度行人目标。又由于采用了跨尺度特征聚合模块提高网络多尺度特征融合能力，对目标

存在部分遮挡时也具有较好的检测效果，实现了多尺度目标的准确检测。从第2列图像可以看出原始图像中的行人目标存在明显的遮挡现象，4种对比算法未能检测出遮挡区域行人目标，而本文算法通过构建多尺度空间金字塔池化层有效融合了目标的多尺度局部区域特征信息，改善了模型因目标遮挡而造成特征形变的鲁棒性，从而实现了遮挡区域的目标检测。第3列图像中由于背景中存在树干等多个高亮伪目标，导致对比算法在检测时出现较多误检现象，本文算法通过建立分层注意力映射模块，在增强行人特征信息的同时抑制背景中的干扰因素从而避免了伪目标因素对检测结果的影响。同样，在第4列图像中几种对比算法将亮度及形状相似的目标误检为行人，而本文算法采用的注意力映射模块使特征提取聚焦于行人目标避免了误检。对比可得本文算法可以在各种复杂环境下准确检测出行人目标，同时在整体上提升正样本的置信度。

为客观评价本文算法的检测性能，选取准确率-召回率(Precision-Recall，P-R)曲线和F1-Score值作为性能评价指标，分别对不同改进网络的检测性能进行定量分析。P-R曲线包括准确率(precision)和召回率(recall)两个指标，其表达式为

$$P_{\text{precision}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}} \tag{10}$$

$$P_{\text{recall}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}} \tag{11}$$

式中，$N_{\text{TP}}$表示正样本被检测正确的个数；$N_{\text{FP}}$表示负样本被检测为正样本的个数；$N_{\text{FN}}$表示正样本被检测为负样本的个数，在不同的置信度水平下计算对应的准确率和召回率获得P-R曲线。通常情况下检测的准确率和召回率相互制约，$F_1$值是两者的调和均值，定义为

$$F_1=2\frac{P_{\text{precision}}\cdot P_{\text{recall}}}{P_{\text{precision}}+P_{\text{recall}}} \tag{12}$$

本文所提出的几种不同改进网络与YOLOv4算法在测试集上的P-R曲线对比如图11所示。可以看出在不同置信度水平下，YOLOv4算法的准确率随Recall的增加出现了明显下降，几种改进网络的准确率下降速度相对缓慢。对P-R曲线上的准确率求平均得到平均准确率(Average Precison，AP)，在图中表现为曲线以下区域面积，几种改进算法的AP均高于YOLOv4算法。

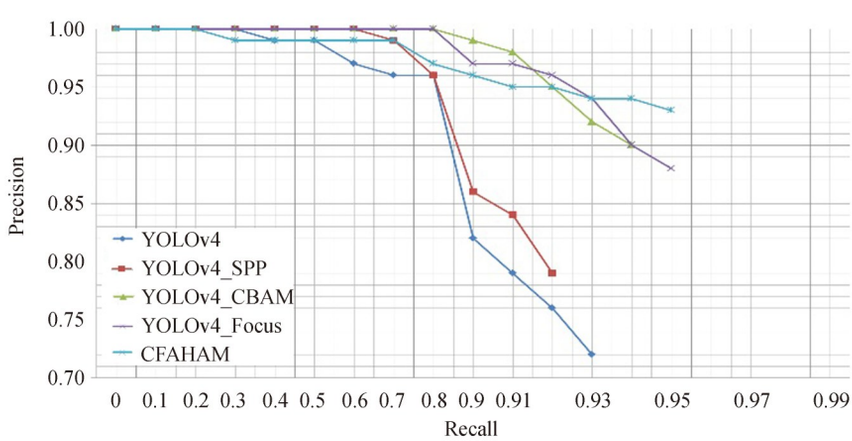



图11 不同算法P-R曲线对比

Fig. 11 P-R curve comparison of different algorithms

图12为本文所提出的CFAHAM检测模型在置信度阈值范围(0～1)内的精度、召回率和F1值的性能曲线。可以看出，在(0.1～0.8)的阈值范围内，几种性能指标可以保持在81%以上。

为了验证本文算法的优势，在OSU Color-Thermal Database、Terravic Motion IR Database和OSU Thermal Pedestrian Database数据集上随机选取了465张图像进行测试，所有实验均采用相同的数据集和参数设置，检测结果如表3所示。

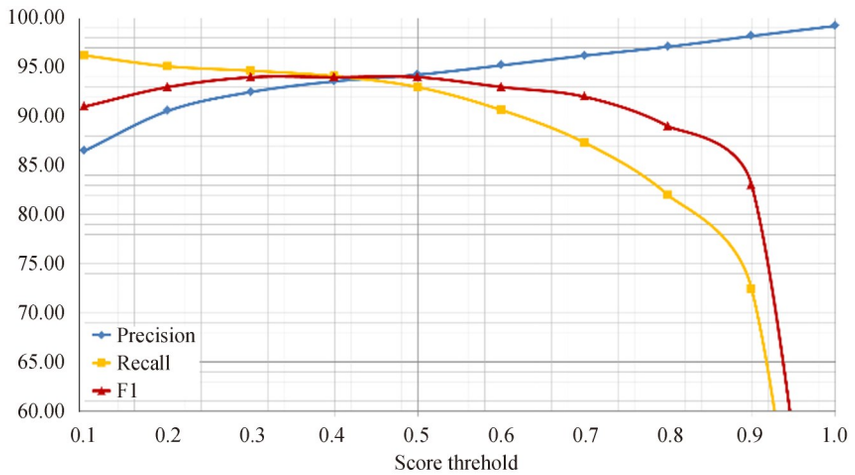



图 12　CFAHAM 算法性能曲线

Fig. 12　Performance curve of the CFAHAM algorithm

**表 3　不同检测算法的对比实验结果**

**Table 3　Comparative experimental results of different detection algorithms**

| Network | AP | Precision | Recall | F1 | Time/s |
|---|---|---|---|---|---|
| SSD | 74.16% | 76.21% | 66.30% | 0.71 | 0.028 2 |
| Faster-RCNN | 83.26% | 73.75% | 85.65% | 0.79 | 0.075 1 |
| YOLOv3 | 84.25% | 81.56% | 81.65% | 0.82 | 0.049 8 |
| YOLOv4 | 87.46% | 91.79% | 77.09% | 0.84 | 0.031 3 |
| **Ours** | **95.37%** | **94.25%** | **92.99%** | **0.94** | **0.035 2** |

## 2.4　鲁棒性验证

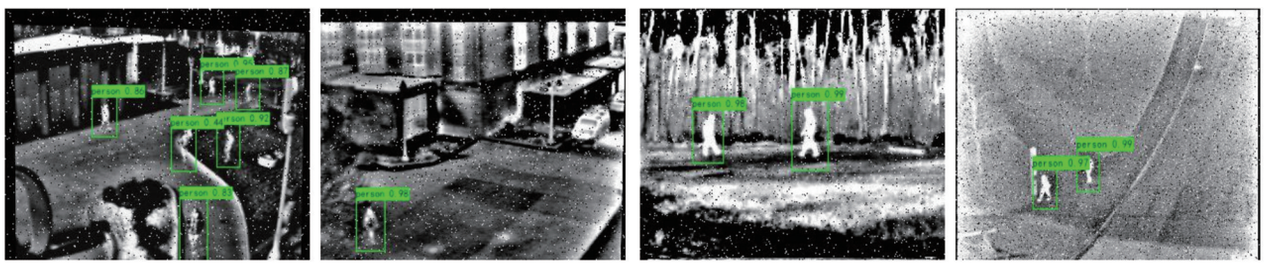
考虑到实际环境中红外图像可能受到噪声干扰以及相机拍摄角度影响使得人体部分处于相机视野中，为了测试本文算法对上述两种情况的有效性分别进行了噪声条件下行人检测实验和处于遮挡区域的行人检测实验。

### 2.4.1　噪声条件

在测试集图像中加入噪声比例为 0.02 的椒盐噪声以验证不同算法在复杂环境下的红外行人检测鲁棒性，本文算法检测结果如图 13 所示。可以看出，当存在噪声时，本文算法依然可以实现复杂背景中行人目标的准确检测。

噪声条件下各种对比算法的检测结果如表 4 所示。可以看出，加入噪声后几种算法的检测结果均有所下降，本文算法虽然检测精度有小幅度降低，但其检测准确率和召回率相比于其他算法依然最高，从而证明其有较好的抗噪能力。

(a) Original image



(b) Detection results after adding noise

图 13 噪声条件下的检测实验

Fig. 13 Detection experiment under noise condition

**表 4 加入噪声后不同算法检测结果对比**

**Table 4 Comparative experimental results of different detection algorithms after adding noise**

| Network | AP | Precision | Recall | F1 |
|---|---|---|---|---|
| SSD | 67.59% | 68.53% | 61.51% | 0.65 |
| Faster-RCNN | 80.03% | 68.58% | 83.76% | 0.75 |
| YOLOv4 | 85.56% | 78.91% | 76.97% | 0.78 |
| **Ours** | **94.52%** | **91.70%** | **89.77%** | **0.91** |

2.4.2 遮挡条件

为了测试本文算法对于人体部分遮挡时的检测效果，从OTCBVS红外行人数据集中选取300张图像进行测试，其中包含382个存在不同程度遮挡的行人目标。检测结果如图14所示，图中存在遮挡的行人目标用红色虚线框标注。可以看出，本文算法可以较好地检测出处于遮挡条件下的人体目标，检测的平均置信度为0.660，召回率为0.945。

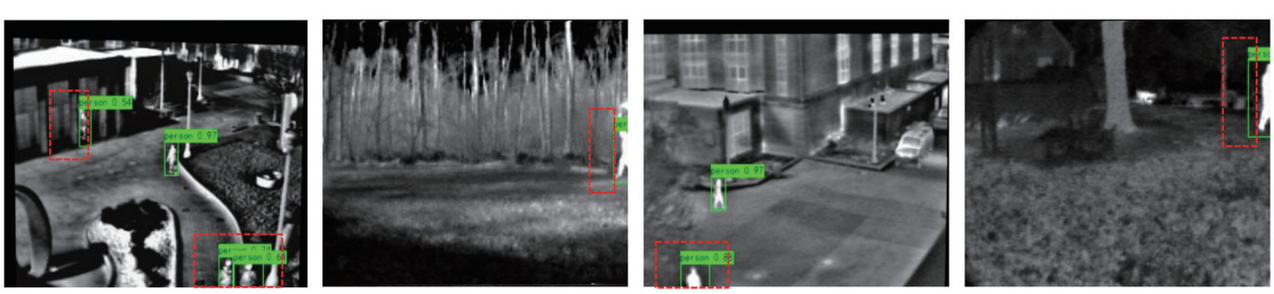


图 14 遮挡条件下的检测结果

Fig. 14 Detection results under occlusion conditions

## 3 结论

本文提出了一种基于跨尺度特征聚合与分层注意力映射的多尺度红外行人检测方法。该方法可以有效解决因红外图像纹理特征弱、空间分辨率差以及存在多尺度、部分遮挡等复杂环境干扰所导致传统算法难以准确检测的问题。

将主干特征提取网络的CBM+CSP_1结构替换为Focus+CBM结构可以有效减少小尺度目标特征的信息丢失；通过构建跨尺度特征聚合模块实现了不同尺度目标特征的有效融合，从而改善了多尺度及部分遮挡区域的行人目标检测性能；通过构建分层注意力映射模块，增强了行人目标在复杂背景中的显著性，较好地解决了行人目标在复杂环境下由于特征表达能力不足而造成的漏检和误检。

在多个公共红外行人检测数据集上与4种经典目标检测算法进行对比实验，证明了本文所提出的红外图像行人检测算法在各种复杂环境下具有较好的检测效果，同时具有良好的实时性和鲁棒性。然而，本文算法在对部分严重遮挡区域的行人目标检测时依然存在不同程度的漏检，后期将针对该问题展开进一步研究。

## 参考文献


[1] WEI Shuigen, WANG Chengwei, CHEN Zhen, et al. Infrared dim target detection based on human visual mechanism[J]. Acta Photonica Sinica, 2021, 50(1): 0110001.
危水根,王程伟,陈震,等. 基于视觉注意机制的红外弱小目标检测[J]. 光子学报,2021, 50(1): 0110001.

[2] JIAO Yifan, YAO Hantao, XU Changsheng. SAN: selective alignment network for cross-domain pedestrian detection[J]. IEEE Transactions on Image Processing, 2021, 30: 2155-2167.

[3] LIU Tianshan, LAM K M, ZHAO Rui, et al. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 315-329.

[4] SONG Zizhuang, YANG Jiawei, ZHANG Dongfang, et al. Low-altitude sea surface infrared object detection based on unsupervised domain adaptation[J]. Acta Optica Sinica, 2022, 42(4): 0415001.
宋子壮, 杨嘉伟, 张东方, 等. 红外探测器间无监督域适应目标检测[J]. 光学学报, 2022, 42(4): 0415001.

[5] WU Shuangchen, ZUO Zhengrong. Small target detection in infrared images using deep convolutional neural networks[J]. Journal of Infrared and Millimeter Waves, 2019, 38(3): 371-380.
吴双忱, 左峥嵘. 基于深度卷积神经网络的红外小目标检测[J]. 红外与毫米波学报, 2019, 38(3): 371-380.

[6] CHEUNG W, HAMARNEH G. N-SIFT: n-dimensional scale invariant feature transform[J]. IEEE Transactions on Image Processing, 2009, 18(9): 2012-2021.

[7] DALAI N, TRIGGS B. Histograms of oriented gradients for human detection[C]. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), June. 20-25, 2005, San Diego, CA, USA: IEEE, 2005: 886-893.

[8] ZHANG C J, LIU J, LIANG C, et al. Image classification using Harr-like transformation of local features with coding residuals[J]. Signal Processing, 2013, 93(8): 2111-2118.

[9] SAID Y, ATRI M, TOURKI R. Human detection based on integral histograms of oriented gradients and SVM[C]. 2011 International Conference on Communications Computing and Control Applications, March. 3-5, 2011, Hammamet, Tunisia: IEEE, 2011: 1-5.

[10] BEGARD J, ALLEZARD N, SAYD P. Real-time human detection in urban scenes: local descriptors and classifiers selection with adaboost-like algorithms[C]. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, June. 23-28, 2008, Anchorage, AK, USA: IEEE, 2008: 1-8.

[11] BISWAS S K, MILANFAR P. Linear support tensor machine with LSK Channels: pedestrian detection in thermal infrared images[J]. IEEE Transactions on Image Processing, 2017, 26(9): 4229-4242.

[12] HIRANMAI M, NIRANJANA K B, NAGARAJ H K. Comparative study of various feature extraction techniques for pedestrian detection[J]. Procedia Computer Science, 2019, 154: 622-628.

[13] LIU Yande, ZENG Tiwei, CHEN Dongbin, et al. Pedestrian detection based on cascade two stage classification[J]. Electronic Measurement Technology, 2018, 41(19): 1-6.

[14] LIU Junming, MENG Weihua. Infrared small target detection based on fully convolutional neural network and visual saliency[J]. Acta Photonica Sinica, 2020, 49(7): 0710003.
刘俊明, 孟卫华. 融合全卷积神经网络和视觉显著性的红外小目标检测[J]. 光子学报, 2020, 49(7): 0710003.

[15] WANG Hongbin, XIAO Song, QU Jiahui, et al. Pansharpening based on multi-branch CNN[J]. Acta Optica Sinica, 2021, 41(7): 0710001.
王洪斌, 肖嵩, 曲家慧, 等. 基于多分支CNN的高光谱与全色影像融合处理[J]. 光学学报, 2021, 41(7): 0710001.

[16] ZHU C, HE Y, SAVVIDES M. Feature selective anchor-free module for single-shot object detection[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June. 15-20, 2019, Long Beach, CA, USA: IEEE, 2020: 840-849.

[17] TIAN Z, SHEN C, CHENG H. FCOS: fully convolutional one-stage object detection[C]. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Nov. 27, 2019, Seoul, Korea (South), IEEE, 2020: 9626-9635.

[18] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. 2014 IEEE Conference on Computer Vision and Pattern Recognition, June. 23-28, 2014: Columbus, OH, USA: IEEE, 2014: 580-587.

[19] GIRSHICK R. Fast R-CNN[C]. 2015 IEEE International Conference on Computer Vision (ICCV), Dec. 7-13, 2015, Santiago, Chile: IEEE, 2015: 1440-1448.

[20] REN S Q, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[21] LI Jianan, LIANG Xiaodan, SHEN Shengmei, et al. Scale-aware fast R-CNN for pedestrian detection[J]. IEEE Transactions on Multimedia, 2018, 20(4): 985-996.

[22] ZHANG Liliang, LIN liang, LIANG X. Is faster R-CNN doing well for pedestrian detection[C]. European Conference

on Computer Vision(2016), Sept. 17, 2016, Cham: Springer, 2016: 443-457.
[23] LIU W, ANGUELOV D, ERHAN D. SSD: single shot multibox detector[C]. 2016 European Conference on Computer Vision(ECCV), Oct. 11-14, 2016, Netherlands: Springer, 2016: 21-37.
[24] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), June. 27-30, 2016, Las Vegas, NV, USA: IEEE, 2016: 779-788.
[25] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July. 21-26, 2017, Honolulu, HI, New York: IEEE, 2017: 6517-6525.
[26] REDMON J, FARHADI A. YOLOv3: An incremental improvement[EB/OL]. (2018-09-30). https://arxiv.org/abs/1804.02767.
[27] BOCHKOVSKIY A, WANG C Y, LIAO H Y. YOLOv4: optimal speed and a accuracy of object detection[J/OL]. (2020-04-23). https://arxiv.org/abs/2004.10934.
[28] ZHAO Bin, WANG Chunping, FU Qiang, et al. Multi-scale infrared pedestrian detection based on deep Attention Mechanism[J]. Acta Optica Sinica, 2020, 40(5): 0504001.
赵斌, 王春平, 付强, 等. 基于深度注意力机制的多尺度红外行人检测[J]. 光学学报, 2020, 40(5): 0504001.
[29] WEN B Y, WU M Q. Study on pedestrian detection based on an improved YOLOv4 algorithm[C]. 2020 IEEE 6th International Conference on Computer and Communications(ICCC), Dec. 11-14, 2020, Chengdu, China: IEEE, 2020: 1198-1202.
[30] DU S J, ZHANG P, ZHANG B F, et al. Weak and occluded vehicle detection in complex infrared environment based on improved YOLOv4[J]. IEEE Access, 2021, 9(9): 25671-25680.
[31] WOO S, PARK J, LEE J, CBAM: convolutional block attention module[C]. European Conference on Computer Vision (ECCV), Sep. 9-8, 2018, Cham: Springer, 2018: 3-19.
[32] DAVIS J W, SHARMA V. Background-subtraction using contour-based fusion of thermal and visible imagery[J]. Computer Vision & Image Understanding, 2007, 106(2-3): 162-182.
[33] DAVIS J W, KECK M A. A two-stage template approach to person detection in thermal imagery[C]. 2005 Seventh IEEE Workshops on Applications of Computer Vision (WACV/MOTION'05) - Volume 1, Jan. 5-7, 2005, Breckenridge, CO, USA: IEEE, 2005: 364-369.

# Infrared Pedestrian Detection Based on Cross-scale Feature Aggregation and Hierarchical Attention Mapping

HAO Shuai[1], GAO Shan[1], MA Xu[1], AN Beiyi[1], HE Tian[1], WEN Hu[2], WANG Feng[3]
(1 *College of Electrical and Control Engineering, Xi'an University of Science and Technology, Xi'an* 710054, *China*)
(2 *College of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an* 710054, *China*)
(3 *College of Physics and Electrical Engineering, Weinan Normal University, Weinan, Shaanxi* 714000, *China*)

**Abstract**: The detection system based on infrared thermal imaging has been extensively used in pedestrian detection because of its strong anti-interference ability, long detection distance and less affected by light and climate change. However, due to its unique thermal radiation imaging, infrared images usually have the defects of unclear texture features and low spatial resolution. At the same time, infrared pedestrian features are easy to be submerged by the bright background, which makes the detection algorithm difficult to locate the object region accurately. In addition, the multi-scale characteristics and mutual occlusion of pedestrian objects also pose a serious challenge to the performance of the detection algorithm. Therefore, aiming at the problem that traditional pedestrian detection algorithms are difficult to detect accurately owing to multi-scale, partial occlusion and environmental interference in infrared pedestrian images, an infrared pedestrian detection algorithm based on cross-scale feature aggregation and hierarchical attention mapping is proposed. Firstly, the CSPdarknet53 structure is utilized as the backbone feature extraction network. On this basis, to reduce the loss of small-scale object feature information during the down-sampling process in the backbone network, the focus module is introduced and added at the input to replace the first residual layer. Using slice segmentation sampling, the spatial dimension information in the original image is extracted to the channel dimension to realize lossless down-sampling. Secondly, to improve the multi-scale feature aggregation ability of the detection network and improve detection accuracy of the network, a cross-

scale feature aggregation module is constructed to integrate the global features and multi-scale local features output by different residual layers of the backbone network. Then, aiming at the problem that infrared images are vulnerable to the effects of self-imaging mechanism and complex background and cannot effectively express pedestrian object features, a hierarchical attention mapping module is constructed by embedding visual attention mechanism into multi-layer feature transfer branches of feature pyramid. In the constructed detection network, the attention mechanisms based on the location, appearance and semantic features of pedestrian objects are established respectively. It establishes semantic and localization associations with spatial and channel dimensions and adaptively adjusts weight coefficients of regions of interest at different scales. The detector can quickly focus on pedestrian objects in the feature extraction process and effectively improve pedestrian detection performance in a complex environment. The ablation experiment proves that the proposed cross-scale feature aggregation module can effectively fuse the features of different scales and improve the pedestrian object detection performance in multi-scale and partially occlusion regions. The constructed hierarchical attention mapping module can enhance the salience of pedestrian objects in the complex background and solve the missed and false detection caused by the lack of feature expressive ability of pedestrian objects in the complex environment. Finally, in order to verify the effectiveness of the proposed algorithm, three infrared pedestrian detection datasets were selected from the OTCBVS common benchmark database for testing. The selected test set covers a variety of complex detection environments, including multi-scale pedestrian objects, highlighted pseudo-objects, fuzzy scenes, etc. The selected experimental scene covers the real pedestrian detection scene well, which can well demonstrate the detection effect of the algorithm in the real scene. In order to verify the advantages of the proposed algorithm, four mainstream object detection algorithms are selected and compared with the proposed algorithm from subjective evaluation and objective evaluation indexes respectively. Experimental results demonstrate that the proposed algorithm has obvious advantages over the contrast algorithm in both subjective and objective evaluation. A large number of experimental results also show that the algorithm can achieve accurate detection of infrared multi-scale pedestrians in a complex environment, with an average accuracy of 95.37% and recall rate of 92.99%.

**Key words**: Infrared pedestrian detection; Multi-scale; Focus module; Spatial pyramid; Attention mechanism

**OCIS Codes**: 100.4996; 040.3060; 040.1880

**Foundation item**: National Natural Science Foundation of China (No. 51804250), China Postdoctoral Science Foundation (Nos.2019M653874XB, 2020M683522), Natural Science Basic Research Program of Shaanxi (Nos.2021JQ-572, 2020JQ-757), Scientific Research Program of Shaanxi Provincial Department of Education (No. 18JK0512), Weinan Science and Technology Project (No. 2020ZDYF-JCYJ-196), Innovation Capability Support Program of Shaanxi (No. 2020TD-021), Xi′an Beilin District Science and Technology Project (No.GX2116)