激光与光电子学进展 先进成像

基于可变形卷积和多尺度残差注意力的多光谱行人检测

张国立[1,2]，常帅[1,2*]，宋延嵩[1,2]，刘天赐[1,2]

[1]长春理工大学光电工程学院，吉林 长春 130022；

[2]长春理工大学空间光电技术研究所，吉林 长春 130022

摘要 目前多光谱行人检测算法大多对可见光与红外图像融合方法展开研究，但是充分融合多光谱图像所需的参数量巨大，会导致检测速度降低。针对这一问题，提出了一种基于时效性较高的YOLOv5s的多光谱行人检测算法。为了保证算法的检测速度，选用可见光与红外光通道方向上的合并方法作为网络的输入，并通过对传统算法的改进来提升检测精度。首先，用可变形卷积替换部分标准卷积，增强了网络对不规则形状的特征目标的提取能力；其次，用多尺度残差注意力模块替换网络中的空间金字塔池化模块，减弱了背景对行人目标的干扰，提升了检测精度；最后，通过改变连接方式，增加大尺度特征拼接层，提升了网络的检测最小尺度，提升了网络对小目标的检测效果。实验结果表明，改进后的算法在检测速度上有明显优势，并比原算法的mAP@0.5和mAP@0.5:0.95分别提升了5.1和1.9个百分点。

关键词 行人检测；可变形卷积；注意力机制；小目标检测；YOLOv5s

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP232131

Multi-spectral Pedestrian Detection Based on Deformable Convolution and Multi-Scale Residual Attention

Zhang Guoli[1,2], Chang Shuai[1,2*], Song Yansong[1,2], Liu Tianci[1,2]

[1]*College of Opto-Electronic Engineering, Changchun University of Science and Technology, Changchun* 130022, *Jilin, China*;

[2]*Institute of Space Photoelectric Technology, Changchun University of Science and Technology, Changchun* 130022, *Jilin, China*

Abstract At present, most of the multi-spectral pedestrian detection algorithms focus on the fusion methods of visible light and infrared images, but the number of parameters to fully fuse multi-spectral images is huge, resulting in lower detection speed. To solve this problem, we propose a multi-spectral pedestrian detection algorithm based on YOLOv5s with high timeliness. To ensure the detection speed of the algorithm, we select the merging method of visible light and infrared light channel direction as the input of the network, and improve the detection accuracy by improving the traditional algorithm. First, some standard convolution is replaced by deformable convolution to enhance the ability of the network to extract irregular shape feature objects. Second, the spatial pyramid pooling module in the network is replaced by multi-scale residual attention module, which weakens the interference of the background to the pedestrian target and improves the detection accuracy. Finally, by changing the connection mode and adding the large-scale feature splicing layer, the minimum detection scale of the network is increased, and the detection effect of the network for small targets is improved. Experimental results show that the improved algorithm has obvious advantages in detection speed, and improves the mAP@0.5 and mAP@0.5:0.95 by 5.1 and 1.9 percentage points over the original algorithm, respectively.

Key words pedestrian detection; deformable convolution; attention mechanism; small target detection; YOLOv5s

收稿日期：2023-09-15；**修回日期**：2023-10-14；**录用日期**：2023-10-20；**网络首发日期**：2023-10-25

基金项目：吉林省教育厅项目(JJKH20200753KJ)；中央引导地方科技发展资金(YDZJ202301ZYTS407)

通信作者：*changshuai@cust.edu.cn

1 引 言

目标检测指从图像或视频等视觉数据中识别出目标，同时指出它们的位置和大小，其在人脸识别[1]、智慧交通[2]、工业检测[3-4]和行人检测[5-6]等方面发挥着越来越重要的作用。其中，行人检测技术在汽车智能驾驶、行人视频监控、机器人传感等领域有重要的研究意义。

传统的基于可见光的行人检测算法易受照明度的影响，在低照明度的情况下算法的检测准确率不高。红外热成像技术是一种被动红外夜视技术，红外图像抗干扰能力强，常常被用来弥补可见光图像低照明度导致的算法准确率不高的缺陷。因此，基于可见光与红外图像融合的多光谱行人检测算法受到了学者们的广泛关注。

Hwang等[7]基于增强热图像的定向梯度直方图特征，提出了多光谱聚合通道特征(ACF)[8]和用于训练和测试的KAIST多光谱行人数据集。结果表明，多光谱算法的检测性能明显优于单光谱算法。在建立了KAIST多光谱行人数据集后，除了研究卷积神经网络(CNN)来提取特征外，许多学者进一步研究了最优的多光谱融合模型。Wagner等[9]提出了一种基于R-CNN(region-CNN)[10]的多光谱行人检测方法，并讨论了多光谱融合模型的优缺点。Liu等[11]将两个基于CNN的子网络进行融合，提出了4种不同的网络结构，并对比得出了最优的融合阶段。Li等[12]提出了光照权重分配机制，通过光照强度来给两种模态加权再进行融合。

上述算法虽然准确率高，但都是基于二阶段算法进行的，所以其处理速度很慢，难以满足行人检测的时效性要求。因此，以YOLO系列算法[13]为基础的一阶段算法被提出。Hsia等[14]提出了一种深度学习框架，以改善光源混淆问题，通过多光谱融合提取高度差异化的多模态特征，利用双流多光谱网络行人检测方法设计了基于MFDs-YOLO(multispectral fusion and double-stream detector-YOLO)信息的多光谱融合双流检测器。方康等[15]设计了一种基于YOLOX的无锚框检查算法，将多模态特征提取解耦为特性特征提取和共性特征提取两阶段。以上算法选取了检测效率较高的YOLO系列算法对多光谱图像融合的方法进行改进，降低了行人检测的漏检率，在一定程度上提升了检测精度。但是由于充分融合多光谱图像所需的参数量大，因此算法的检测速度有所降低，难以满足行人检测所需的检测效率。

针对上述问题，本文提出了一种基于改进YOLOv5s算法的多光谱行人检测方法。在众多融合方法中选择早期融合，即合并可见光图像和红外图像，以4通道图像作为网络的输入端，在不改变网络结构的前提下，消除了图像融合所带来的参数量，保证了检测速度。改进后的算法在检测精度上也有所提高。

2 改进的YOLOv5s算法

YOLOv5网络共有4种网络模型结构，分别是YOLOv5x、YOLOv5s、YOLOv5m和YOLOv5l，其中YOLOv5s网络模型最小，检测速度最快。由于对实时性要求较高，因此选择YOLOv5s网络模型进行改进。

2.1 可变形卷积

CNN对大尺度多形变目标的建模存在固有的缺陷，因为CNN只对输入特征图的固定位置进行采样。在同一层特征图中，所有特征点的感受野是一样的，但由于不同位置可能对应着不同尺度或形状的物体，因此需要对尺度或者感受野大小进行自动调整。

针对此问题，引入可变形卷积(DCN)[16]，自适应地调节尺度特征图的尺度和感受野。在YOLOv5s算法的主干网络(backbone)中用DCN替换掉前两层的传统卷积，以提高模型对形变目标的建模能力。传统卷积和DCN的采样位置的对比如图1所示。

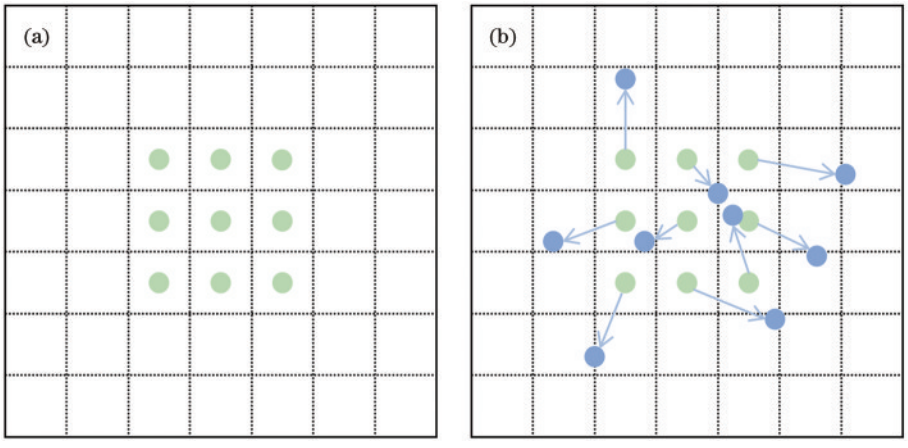

图1 采样位置。(a)常规卷积；(b) DCN

Fig. 1 Sampling position. (a) Regular convolution; (b) DCN

常规的卷积操作主要分为两步：1)在输入的特征图上使用规则网格 R 进行采样；2)使用卷积核 k 对采样点进行加权运算。

定义一个尺寸为3×3、扩张率为1的卷积核表示为

$$k=\{(-1,-1)(-1,0),\cdots,(1,1)\},\tag{1}$$

可以计算出输出特征图 $y(p_0)$ 上的每个位置 p_0，计算公式为

$$y(p_0)=\sum_{p_0\in R}w(p_n)\times x(p_0+p_n),\tag{2}$$

式中：R 定义了感受野的大小和扩张；p_n 为卷积核元素相对于卷积核中心的偏移量；$w(p_n)$ 表示第 n 个位置的点所对应的权重；$x(p_0+p_n)$ 表示第 n 个位置的点的像素值。对于3×3尺寸的卷积核，左上角位置像素值为(−1, −1)，右下角位置像素值为(1, 1)。

在DCN的操作中，通过给规则网格 R 增加一个偏移量 $\{\Delta p_n|n=1,2,\cdots,N\}$，$N=|R|$ 进行扩张，那么同样的位置 p_0 的输出特征图 $y(p_0)$ 变为

$$y(p_0)=\sum_{p_0\in R}w(p_n)\times x(p_0+p_n+\Delta p_n)。\tag{3}$$

由于偏移量 Δp_n 通常是小数，因此需要通过双线性插值法计算 x 的值，表示为

$$x(p)=\sum_q G(q,p)\times x(q),\tag{4}$$

式中：p 为DCN对应区域的任意位置；q 为特征图 x 中采样点的像素值；$G(q,p)$ 表示一个二维双线性插值。

图2是一个尺寸为3×3的卷积核的DCN过程。

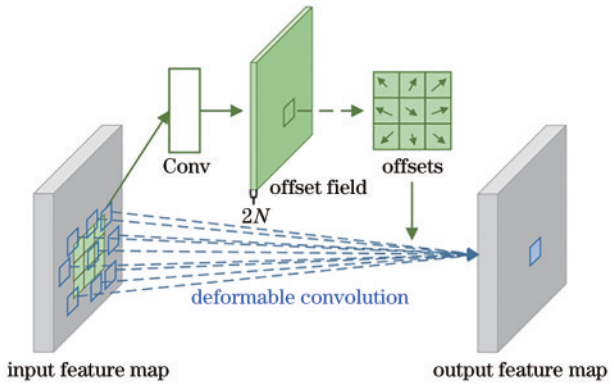

图2　3×3 DCN

Fig. 2　3×3 DCN

2.2　多尺度残差注意力模块

空间金字塔池化(SPP)模块镶嵌在YOLOv5s主干网络最后一个C3模块之前，由4个并行的分支构成，分别是池化核为5×5、9×9、13×13的最大池化操作和一个跳跃连接，如图3所示。不同尺度的最大池化操作和跳跃连接能够让网络学习到图片不同尺度的特征。将局部和全局特征融合，就丰富了特征图的表达能力。最大池化操作虽然能减少冗余信息，但是也容易造成特征信息的丢失。

针对此问题，仿照注意力机制模块[17]，借鉴残差网络[18]，设计了一种多尺度残差注意力模块，以减少最大池化操作导致的特征信息丢失。注意力机制是对输入进行加权再输出，通过对图片的加权处理凸显重要的区域，同时弱化无效特征信息。而残差网络则可以解决增加网络深度带来的网络退化问题，确保通过增加网络深度能够提高网络性能。多尺度残差注意力模块的结构如图4所示。

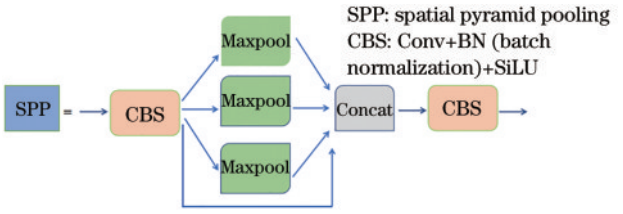

图3　SPP模块示意图

Fig. 3　Schematic diagram of SPP module

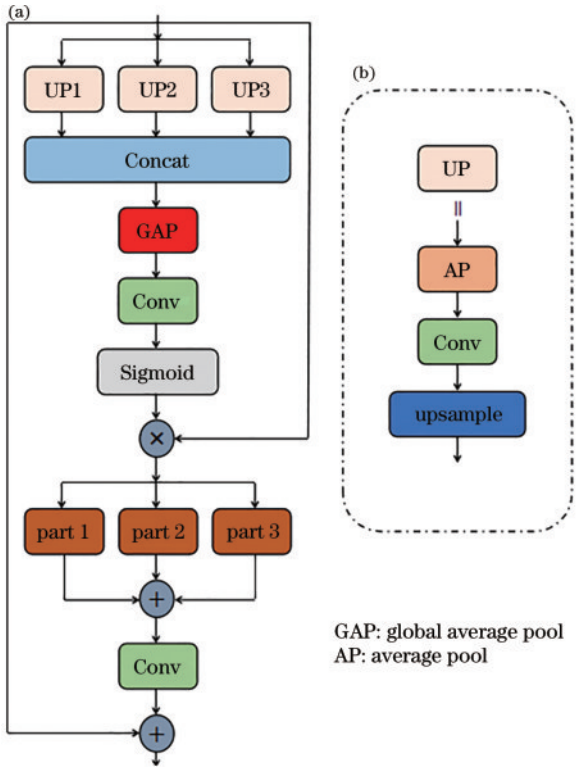

图4　多尺度残差注意力模块示意图。(a)整体结构；(b) UP模块结构

Fig. 4　Schematic diagram of multi-scale residual attention module. (a) Overall structure; (b) structure of UP mudule

多尺度残差注意力模块主要由平均池化(AP)层、上采样层、全局平均池化(GAP)层、Sigmoid函数和残差模块组成。输入特征图首先经过三个不同的UP(AP+Conv+upsample)操作[UP结构如图4(b)所示]，即先经过一个AP层，再经过一个卷积层后进行上采样操作。每个UP中AP的卷积核的尺寸不同，分别是5×5、10×10、16×16，这样可以保证在不同的感受野下对特征图进行提取，减少图片中有效特征的损失。将经过上采样操作的特征图进行Concat拼接，提取不同尺度的图片特征。随后经过一个GAP层，其作

用是减少输入特征图的参数量，将特征图的每个通道都赋予一个值，为后面的加权操作作铺垫。接着，通过Sigmoid激活函数给各个通道赋予不同的权重，并将原始特征图与相应的权重相乘。这部分结构仿照了注意力机制，实现了给特征图加权的效果。此外，用AP代替最大池化的方法，在保证多尺度学习特征图的同时，减少了有效特征信息的丢失。

由于开始的输入是由Concat进行拼接的，因此拼接后的通道数是原始输入的三倍，这增加了网络模型的计算参数量。将加权处理后的特征图按照通道方向进行拆分，然后进行相加操作，使输出的通道数与原始输入保持一致，消除增加参数量带来的弊端。最后，输出特征图经过一层卷积后与输入特征图恒等映射支路进行相加，此处借鉴了残差网络的结构。

2.3　大尺度特征拼接层

连接网络（neck）部分采取了特征金字塔网络（FPN）[19]和路径聚合网络（PAN）[20]相结合的结构，如图5所示。FPN是自上而下的结构，将高层特征通过上采样的方法与低层特征融合；PAN是自下而上的结构，将低层特征通过下采样的方法与高层特征融合。FPN+PAN的结构实现了多尺度的特征融合。融合后的特征层输入检测端进行最终检测。

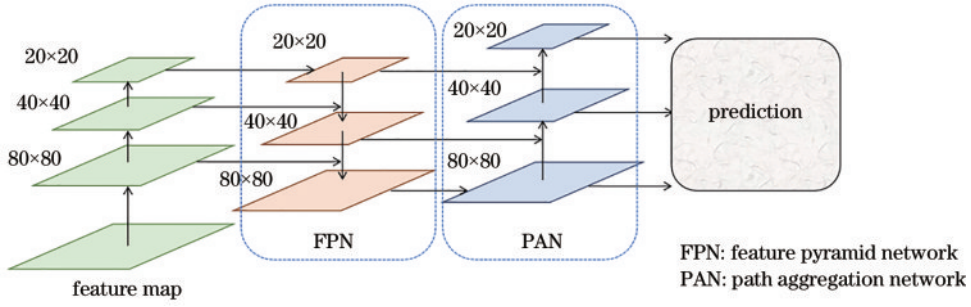

图5　改进前的连接网络

Fig. 5　Neck before improvement

传统的算法只能得到尺度大小为80×80、40×40、20×20的特征图，分别用于检测尺度大小为8×8以上、16×16以上、32×32以上的目标。但是在行人检测的任务中，行人目标在不同背景和角度下呈现出的比例是不同的，而传统的YOLOv5s只能检测最小尺度为8×8的目标，对小目标的检测效果较差。

针对此问题，在neck中加入大尺度特征拼接层，使网络可以检测出尺度大小为4×4的行人目标，提升算法对小目标的检测性能。改进后的连接网络如图6所示。

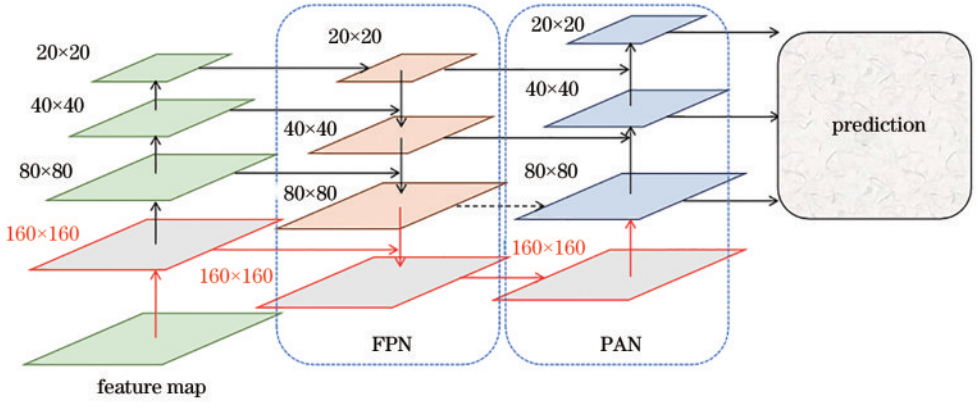

图6　改进后的连接网络

Fig. 6　Neck after improvement

首先，令原neck部分中FPN结构最底层上采样的特征层继续进行上采样的操作，得到尺度大小为160×160的特征层图，同时断开与PAN结构的直接连接；然后，令backbone部分中的第4层（浅层小目标特征层，尺度大小为160×160）与FPN结构中新得到的尺度大小为160×160的特征图进行拼接；最后，利用拼接后的特征图在PAN结构进行下采样。最终在不改变检测头个数的情况下，提升算法的检测效率和对于小目标的检测性能。

改进后的YOLOv5s网络如图7所示。本研究侧重于提升算法的检测速度，将可见光和红外光图像合并作为输入，可见光图像在前红外光图像在后。本网络中：DCN替换掉了backbone部分前两层的标准卷积；backbone部分最后一个C3模块前的SPP模块被多尺度残差注意力模块所代替（图7中多尺度残差注意力机制用“MRA”表示）；大尺度特征拼接层连接方法如图7所示。

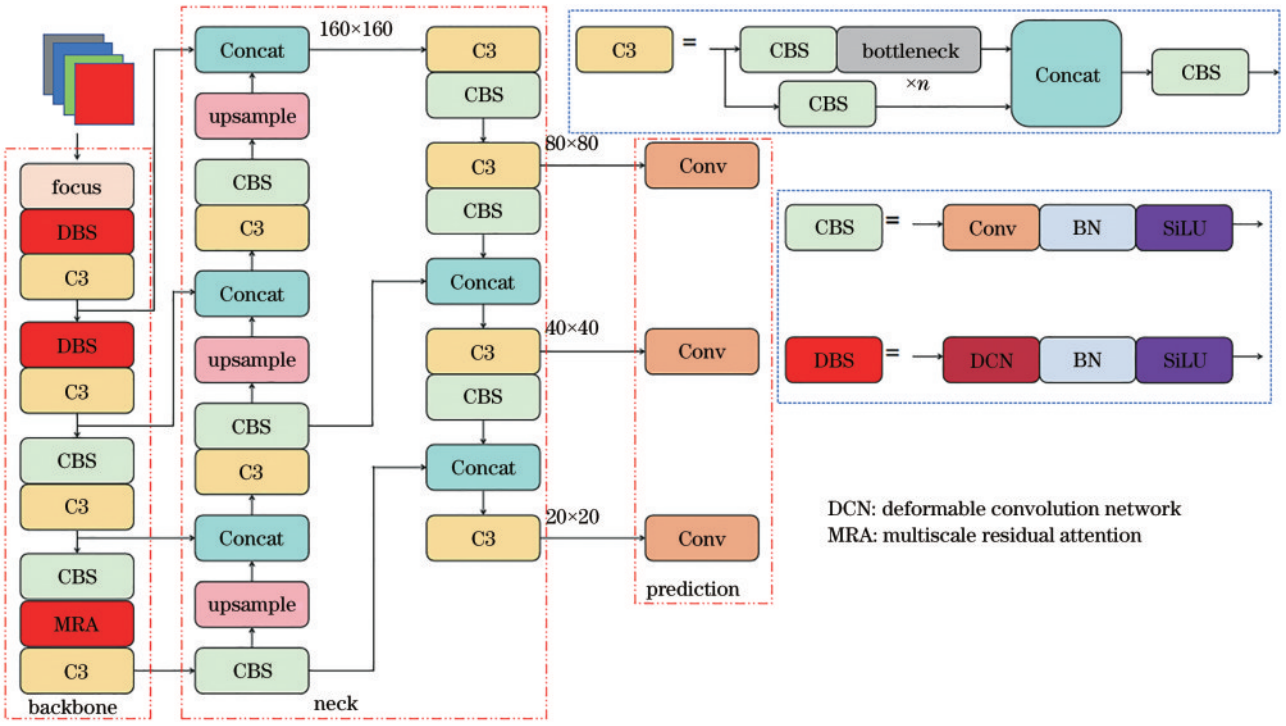

图 7　改进后的 YOLOv5s 网络

Fig. 7　Improved YOLOv5s network

3　分析与讨论

3.1　数据集

实验数据集使用的是 KAIST 行人数据集[7]，该数据集共有 95328 张图片，并且每张图片都对应三通道彩色图像和单通道红外图像，共有 103128 个密集注释。该数据集采集了白天和夜晚的校园、街道和乡村的各式各样的交通路段场景。数据集共有三个类别的标签，分别是 person（比较好分辨的单个行人）、people（不太好分辨的多个行人）和 cyclist（骑自行车的人）。

利用 Liu 等[11]提出的清洗规则对数据集进行清洗：训练集每隔两张图片选取一张，即每三张图片选取一张作为实验训练集的数据，并且除掉其中不含有任何行人样本的图片。选出的图片至少含有一个行人，减少严重遮挡和像素值过小的行人目标。经过清洗后的训练集有 7601 张图片，其中 4755 张是白天场景，2846 张是夜晚场景。测试集每隔 19 张图片选取一张，即每 20 张图片选取一张作为实验测试集的数据，并且保留不含有任何行人的样本图片。经过清洗后的测试集有 2252 张图片，其中 1455 张是白天场景，797 张是夜晚场景。只留下 person 标签，其余均删除。

3.2　实验环境与参数设置

实验环境如表 1 所示。训练过程采用随机梯度下降法，初始学习率为 0.01，动量参数为 0.92，epoch 设置为 300，batch size 设置为 16，权值衰减系数为 0.0005。在模型的初始训练中用 warm-up 方法进行三个 epoch 的预热训练，动量参数为 0.8。原始图片尺寸为 640 pixel×512 pixel，训练时将图片尺寸变换为 640 pixel×640 pixel。

表 1　实验环境配置

Table 1　Experimental environment configuration

Parameter	Experimental environment
Operating system	Windows 11
CPU	Intel Core i5-12500H @ 2.40 GHz
GPU	GeForce RTX 3060
Memory	32 Gb
Python	3.8
Deep learning framework	PyTorch 1.7.0、CUDA 11.1

3.3　评价指标

实验的性能评价指标有：mAP@0.5、mAP@0.50:0.95 和 FPS。其中：mAP 指的是均值平均精度，是对精确率（P）和召回率（R）的一种综合处理指标，表示 P-R 曲线与坐标轴所围成图形的面积。mAP@0.5 表示 IOU（交并比，是预测的边框和真实边框的交集和并集的比值）为 0.5 时的 mAP，mAP@0.50:0.95 表示 IOU 从 0.50 到 0.95（步长为 0.05）时 mAP 的平均值。FPS 表示帧率，是用来衡量算法速度的指标，FPS 值越大表示检测速度越快。

P、R、mAP 的计算公式表示为

$$P=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{P}}},\tag{5}$$

$$R=\frac{T_{\mathrm{P}}}{T_{\mathrm{P}}+F_{\mathrm{N}}},\tag{6}$$

$$V_{\mathrm{mAP}}=\int_0^1 P\mathrm{d}R, \tag{7}$$

式中：T_{P}是正例被正确预测的数量；F_{P}是负例被错误预测为正例的数量；F_{N}是正例被错误预测的数量。

3.4　消融实验

为验证每个改进点对算法的提升效果，设计消融实验如下：如表1所示，“Group”代表组别；“DCN”代表可变形卷积的引入；“MRA”代表将多尺度残差注意力模块替换掉原SPP模块；“Small target”代表大尺度特征拼接层的加入。1组是原始YOLOv5s算法，8组是所提方案，2组至7组则是加入不同改进后的算法。表2中的“√”代表引入该模块。消融实验结果如表2所示。

表2　消融实验结果

Table 2　Ablation experimental results

Group	DCN	MRA	Small target	mAP@0.5 /%	mAP@0.5:0.95 /%	FPS /(frame/s)
1				49.7	26.3	63.8
2	√			51.1	26.9	60.7
3		√		50.6	26.5	61.4
4			√	51.7	27.0	60.3
5	√	√		51.9	27.3	58.7
6	√		√	53.5	27.7	56.4
7		√	√	52.6	27.5	56.8
8	√	√	√	**54.8**	**28.2**	**54.4**

由表2可知：与原算法相比，引入DCN进行特征提取后，mAP@0.5和mAP@0.50:0.95分别提升了1.4和0.6个百分点；在应用改进后的多尺度残差注意力模块后，mAP@0.5和mAP@0.50:0.95分别提升了0.9和0.2个百分点；在加入大尺度特征拼接层后，mAP@0.5和mAP@0.50:0.95分别提升了2.0和0.7个百分点。多项的组合实验中，算法的mAP@0.5和mAP@0.50:0.95也都有相应的提升。由于新增了偏移量、增加了SPP的复杂程度以及新增了检测层，模型的计算量变大，检测速度都略有下降，改进后算法的FPS比原算法低了9.4 frame/s，但仍在实时检测时间的可控范围内。

综上所述，所提三处对算法的改进都可以在一定程度上提升模型的性能，证明了所提方法的有效性和可行性。改进后的算法(group 8)相比原始的YOLOv5s算法，mAP@0.5提升了5.1个百分点，其*P-R*曲线如图8所示；mAP@0.50:0.95提升了1.9个百分点。证明所提方法提升了网络对目标的检测能力。

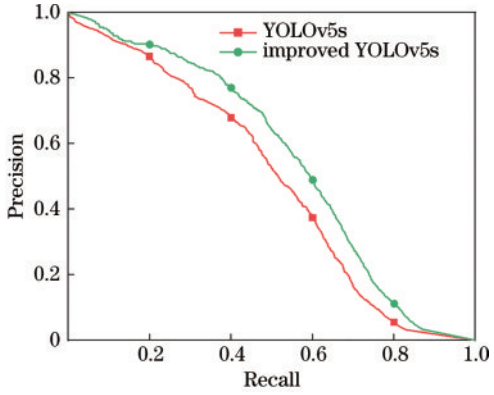

图8　*P-R*曲线

Fig. 8　Curves of *P-R*

3.5　对比实验

为了验证合并可将光图像和红外图像的多光谱方法具有更快的检测速度，选取多种经典融合方法在KAIST数据集上进行实验比较，选取的融合方法有ACF[8]、Halfway Fusion[11]、IAF-RCNN[12]、CIAN[21]、DSMN[14]和MBNet[22]，其对应的平台和检测速度如表3所示。

表3　融合方法对比实验结果

Table 3　Comparative experimental results of fusion methods

Method	Platform	FPS /(frame/s)
ACF	MATLAB	0.37
Halfway Fusion	TITAN X	2.36
IAF-RCNN	TITAN X	4.76
CIAN	1080Ti	14.28
DSMN	3090Ti	1.32
MBNet	1080Ti	14.29
Ours	3060Ti	**54.40**

根据表3的实验结果，所提算法的检测速度明显优于其他算法。由于行人检测任务对实时性的要求较高，证明了所提方法的可行性，验证了合并可见光图像和红外图像的方法在检测速度上存在优势。

为了进一步验证所提改进算法的先进性，以mAP@0.5和FPS作为评价指标，将所提算法与其他经典的目标检测算法在同种环境下进行比较。选取的进行对比实验的算法有Faster R-CNN[23]、SSD(single shot MultiBox detector)[24]、YOLOv4-tiny[25]和YOLOv7[26]。对比实验结果如表4所示。

由表4可知，改进后算法的mAP@0.5比其他算

表 4　不同算法的对比实验结果

Table 4　Comparative experimental results of different methods

Method	mAP@0.5 /%	FPS /(frame/s)
Faster R-CNN	46.7	14.8
SSD	38.6	42.9
YOLOv4-tiny	42.5	53.6
YOLOv5s	49.7	63.8
YOLOv7	51.6	39.3
Ours	**54.8**	**54.4**

法要高，证明了所采取的改进方式的有效性。此外，有着更复杂网络结构的 YOLOv7 算法的 mAP@0.5 比原始的 YOLOv5s 算法要高，比改进后的 YOLOv5s 低，说明了所提改进方案的成功。在处理速度上，改进后的算法的 FPS(54.4 frame/s)虽然较原算法略有降低，但依然优于其他算法。

综上所述，改进后的网络在平均精度上均优于其他算法，在处理速度上也依旧满足行人检测的需求。

3.6　算法验证

将所提算法与原始的 YOLOv5s 算法以及经典的多模态行人检测算法 MBNet 的检测结果进行可视化对比，对比结果如图 9 所示。图 9 分别展示了复杂背景场景、严重遮挡行人目标场景、模糊背景小目标场景以及弱特征行人目标场景下不同算法的检测结果。

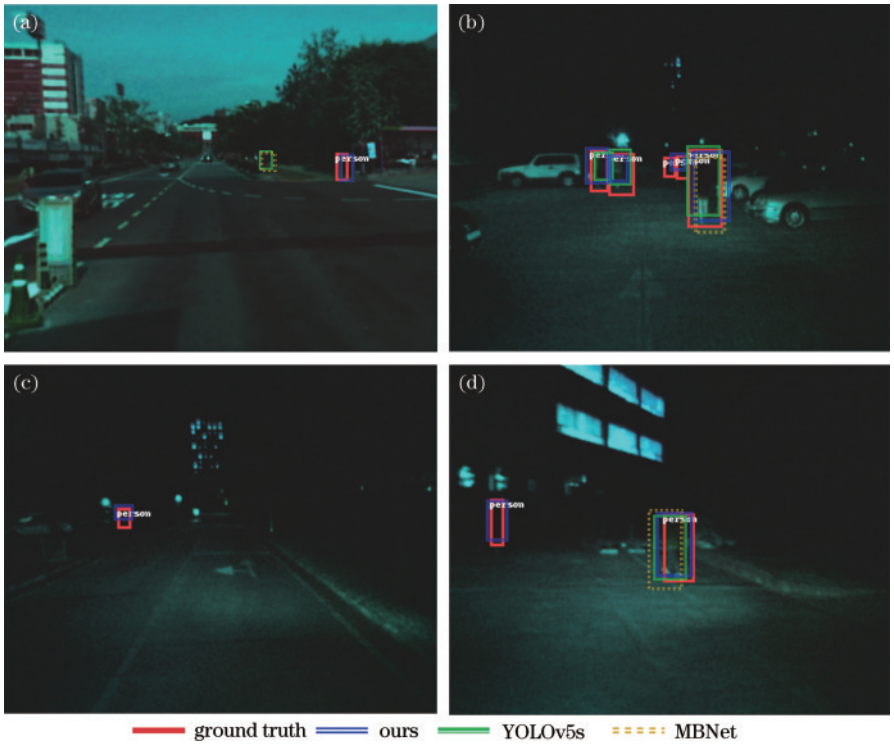

图 9　不同算法检测结果的可视化对比。(a)复杂背景场景；(b)严重遮挡行人目标场景；(c)模糊背景小目标场景；(d)弱特征行人目标场景

Fig. 9　Visualization comparison of detection results by different algorithms. (a) Complex background scene; (b) heavily occlusion pedestrian target scene; (c) blur background small target scene; (d) weak feature pedestrian target scene

通过观察图像检测结果可知，改进后的算法在各个场景下表现良好，可以检测出原算法漏检的行人目标，且尺度较小的行人目标的检测效果也有所改善。

4　结　　论

提出了一种基于改进 YOLOv5s 算法的多光谱行人检测方法。通过引入 DCN，达到了自适应调节特征图尺度和感受野的目的。加入多尺度残差注意力模块保证了特征图可以得到多尺度的学习。通过加权的方式让特征图中的行人目标更加明显。通过加入大尺度特征拼接层使算法可以捕捉较小的目标。最后，以 mAP 和 FPS 为评价指标在 KAIST 数据集上进行了实际测试，并且与其他算法进行了对比。对比结果证明了所提算法在行人检测速度上具有明显优势，有着较高的检测精度，能够很好地满足对实时性要求较高的行人检测任务。

参　考　文　献

[1] Wang Z Y, Huang B J, Wang G C, et al. Masked face recognition dataset and application[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2023, 5(2): 298-304.

[2] Karami Z, Kashef R. Smart transportation planning: data, models, and algorithms[J]. Transportation Engineering, 2020, 2: 100013.

[3] 袁盼, 谭竹嫣, 张旭, 等. 工业气体泄漏红外成像检测及差分光谱滤波检测方法研究[J]. 红外与激光工程,

2022, 51(8): 20210714.
Yuan P, Tan Z Y, Zhang X, et al. Research on infrared imaging detection and differential spectrum filtering detection methods for industrial gas leakage[J]. Infrared and Laser Engineering, 2022, 51(8): 20210714.

[4] 李乾, 薛俊鹏, 张启灿, 等. 利用相机响应曲线实现高反光元件三维面形测量[J]. 光学学报, 2022, 42(7): 0712001.
Li Q, Xue J P, Zhang Q C, et al. Three dimensional shape measurement of high reflective elements using camera response curve[J]. Acta Optica Sinica, 2022, 42(7): 0712001.

[5] 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.
Zhao B, Wang C P, Fu Q, et al. Multi-scale infrared pedestrian detection based on deep attention mechanism [J]. Acta Optica Sinica, 2020, 40(5): 0504001.

[6] 邹梓吟, 盖绍彦, 达飞鹏, 等. 基于注意力机制的遮挡行人检测算法[J]. 光学学报, 2021, 41(15): 1515001.
Zou Z Y, Gai S Y, Da F P, et al. Occluded pedestrian detection algorithm based on attention mechanism[J]. Acta Optica Sinica, 2021, 41(15): 1515001.

[7] Hwang S, Park J, Kim N, et al. Multispectral pedestrian detection: benchmark dataset and baseline[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 1037-1045.

[8] Dollár 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): 1532-1545.

[9] Wagner J, Fischer V, Herman M, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks[C]//24th European Symposium on Artificial Neural Networks, April 27-29, 2016, Bruges, Belgium. [S.l.: s.n.], 2016: 509-514.

[10] Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[11] Liu J J, Zhang S T, Wang S, et al. Multispectral deep neural networks for pedestrian detection[EB/OL]. (2016-11-08)[2023-05-06]. https://arxiv.org/abs/1611.02644.

[12] Li C Y, Song D, Tong R F, et al. Illumination-aware faster R-CNN for robust multispectral pedestrian detection[J]. Pattern Recognition, 2019, 85: 161-171.

[13] 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. New York: IEEE Press, 2016: 779-788.

[14] Hsia C H, Peng H C, Chan H T. All-weather pedestrian detection based on double-stream multispectral network [J]. Electronics, 2023, 12(10): 2312.

[15] 方康, 黄琴, 王克琪, 等. 基于改进YOLOX的多光谱行人检测算法[J/OL]. 小型微型计算机系统: 1-9[2023-05-06]. https://doi.org/10.20009/j.cnki.21-1106/TP.2022-0347.
Fang K, Huang Q, Wang K Q, et al. Multispectral pedestrain detection based on improved YOLOX[J/OL]. Journal of Chinese Computer Systems: 1-9[2023-05-06]. https://doi.org/10.20009/j.cnki.21-1106/TP.2022-0347.

[16] Wang W H, Xie E Z, Song X G, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Republic of Korea. New York: IEEE Press, 2020: 8439-8448.

[17] Dai J F, Qi H Z, Xiong Y W, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 764-773.

[18] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks[EB/OL]. (2014-09-10) [2023-05-04]. https://arxiv.org/abs/1409.3215.

[19] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[20] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[J]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8759-8768.

[21] Zhang L, Liu Z Y, Zhang S F, et al. Cross-modality interactive attention network for multispectral pedestrian detection[J]. Information Fusion, 2019, 50: 20-29.

[22] Zhou K L, Chen L S, Cao X. Improving multispectral pedestrian detection by addressing modality imbalance problems[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12363: 787-803.

[23] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[24] Ren S, He K, Girshick R, et al. Faster R-CNN: toward real-time object detection with region proposal networks [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[25] Jiang Z C, Zhao L Q, Li S Y, et al. Real-time object detection method based on improved YOLOv4-tiny[EB/OL]. (2020-11-09) [2023-05-04]. https://arxiv.org/abs/2011.04244.

[26] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 17-24, 2023, Vancouver, BC, Canada. New York: IEEE Press, 2023: 7464-7475.