激光与光电子学进展 先进成像

基于双模态图像关联式融合的行人实时检测

毕程程[1,2,3]，黄妙华[1,2,3*]，刘若璎[1,2,3]，王量子[1,2,3]

[1]武汉理工大学现代汽车零部件技术湖北省重点实验室，湖北 武汉 430070；

[2]武汉理工大学汽车零部件技术湖北省协同创新中心，湖北 武汉 430070；

[3]武汉理工大学湖北省新能源与智能网联车工程技术研究中心，湖北 武汉 430070

摘要 为解决行人检测任务中低能见度场景下单模态图像漏检率高和现有双模态图像融合检测速度低等问题，提出了一种基于双模态图像关联式融合的轻量级行人检测网络。网络模型基于YOLOv7-Tiny设计，主干网络嵌入关联式融合模块RAMFusion用以提取和聚合双模态图像互补特征；将特征提取部分的1×1卷积替换为带有空间感知能力的坐标卷积；引入Soft-NMS改善结群行人漏检问题；嵌入注意力机制模块来提升模型检测精度。在公开的红外与可见光行人数据集LLVIP上的消融实验表明：与其他融合方法相比，所提方法行人漏检率降低、检测速度显著提高；与YOLOv7-Tiny相比，改进后的模型检测精度提高了2.4%，每秒检测帧数达到124 frame/s，能够满足低能见度行人实时检测需求。

关键词 行人检测；红外与可见光图像；关联式融合；轻量化网络；注意力机制；YOLOv7-Tiny

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP230933

Real-Time Pedestrian Detection Based on Dual-Modal Relevant Image Fusion

Bi Chengcheng[1,2,3], Huang Miaohua[1,2,3*], Liu Ruoying[1,2,3], Wang Liangzi[1,2,3]

[1]*Hubei Key Laboratory of Advanced Technology for Automotive Components, Wuhan University of Technology, Wuhan* 430070, *Hubei, China*;

[2]*Hubei Collaborative Innovation Center for Automotive Components Technology, Wuhan University of Technology, Wuhan* 430070, *Hubei, China*;

[3]*Hubei Research Center for New Energy & Intelligent Connected Vehicle, Wuhan University of Technology, Wuhan* 430070, *Hubei, China*

Abstract In order to solve the problems of high missing detection rate of single-model images and low detection speed of existing dual-model image fusion in pedestrian detection tasks under low visibility scenes, a lightweight pedestrian detection network based on dual-model relevant image fusion is proposed. The network model is designed based on YOLOv7-Tiny, and the backbone network is embedded with RAMFusion, which is used to extract and aggregate dual-model image complementary features. The 1×1 convolution of feature extraction is replaced by coordinate convolution with spatial awareness. Soft-NMS is introduced to improve the pedestrian omission in the cluster. The attention mechanism module is embedded to improve the accuracy of model detection. The ablation experiments in public infrared and visible pedestrian dataset LLVIP show that compared with other fusion methods, the missing detection rate of pedestrians is reduced and the detection speed of the proposed method is significantly increased. Compared with YOLOv7-Tiny, the detection accuracy of the improved model is increased by 2.4%, and the detection frames per second is up to 124 frame/s, which can meet the requirements of real-time pedestrian detection in low-visibility scenes.

Key words pedestrian detection; infrared and visible images; relevant fusion; lightweight network; attention mechanism; YOLOv7-Tiny

1 引 言

行人检测技术在智能交通系统、智能安防监控和智能机器人等领域均表现出极高的应用价值，已经成为计算机视觉领域的重要研究方向之一[1]。但是由于行人所处环境的复杂性，特别是在夜间、大雾等低能见

收稿日期：2023-03-22；**修回日期**：2023-04-10；**录用日期**：2023-04-23；**网络首发日期**：2023-05-03

基金项目：国家重点研发计划(2018YFE0105500)

通信作者：*mh_huang@163.com

度场景下，行人检测方法面临着精度和效率均不足的严峻挑战。

传统的行人检测方法里，Xu 等[2]通过人体特征部分如头部、肩膀对行人进行预测，再利用支持向量机进行分类；Ge 等[3]使用双阈值分割方法分割近距离行人目标。随着信息技术的发展，深度学习方法已经被广泛应用于检测任务。当前检测模型主要分为两类：以 Faster R-卷积神经网络（CNN）[4]、Mask R-CNN[5]等算法为代表的两阶段检测算法和以 SSD（Single Shot MultiBox Detector）[6]、YOLO[7]等算法为代表的单阶段检测算法。两阶段算法检测精度高但实时性远远低于单阶段算法，近年来，YOLO 系列模型被不断优化和改进，检测精度和速度大幅提高，具备了更加优秀的综合检测性能[8]。YOLO 系列模型具备高效的实时性，适用于目标检测的诸多领域，研究者们基于此进行了一些红外与可见光相关的行人检测研究工作。何自芬等[9]提出一种多尺度特征融合红外行人实时检测模型，用轻量化网络替换 YOLOv5 主干网络，检测速度有所提高，但没有利用到可见光的丰富纹理信息；孙颖等[10]提出一种基于 YOLOv5-n 改进的双模态行人检测算法，该方法利用门控融合网络自适应调节红外与可见光特征的权重分配，然而该方法在图像融合阶段单方面提取特征再融合，部分聚合网络没有实现行人特征信息交互；刘子龙等[11]提出一种基于 YOLOv5 改进的利用双模态图像动态权重融合的检测算法，该方法对于道路的小目标和密集目标的检测具有一定优势，但是采用的像素级融合方法不能有效聚合丰富的红外与可见光特征信息，行人检测效果不佳。

上述基于 YOLO 系列模型改进的深度学习方法在行人检测上取得了一定成效，但是在双模态图像融合方面，对红外与可见光特征信息的提取过于泛化，缺乏对检测目标的关注。传统的融合算法主要依赖手动设计特征提取和融合规则如双树复小波变换[12]、剪切波变换[13]等。近年来，基于深度神经网络的图像融合方法快速发展。2019 年 Li 等[14]提出了一种基于自编码器的深度学习框架，编码层采用的密集块（dense block）能够尽可能保留更多的深度特征，训练时将编码层和解码层权重固定，再用适应性融合策略将编码层提取的深层特征融合。Xu 等[15]提出了一种无监督的图像融合深度网络，使用统一的信息度量方法来决定源图像的信息保留程度。Li 等[16]继续设计了一种基于残差结构的多尺度特征融合网络 RFN-Nest，增强了融合效果。这些融合方法能够聚合多模态信息，但网络较为复杂、计算量过大，嵌入检测网络无法满足实时性需求。

为平衡双模态信息融合检测精度和检测速度，本文设计了一种基于双模态关联式融合的轻量级行人实时检测网络模型。所提模型以 YOLOv7-Tiny[17]为基准，主干网络输入端嵌入新设计的关联式融合模块（RAMFusion），通过关联结构和残差结构针对性地提取可见光图像的背景纹理信息和红外图像的行人信息，更加关注行人特征；融合网络模块没有采用较深的网络，参数量少，以提高融合速度；主干网络用坐标卷积（CoordConv）[18]替换 1×1 卷积（Conv），增加网络的空间感知能力，坐标转换平滑，有利于行人特征聚合；嵌入注意力模块以增强网络检测精度。在红外与可见光行人数据集上的实验结果表明了所提模型的优秀检测性能。

2 双模态检测模型

2.1 YOLOv7-Tiny 算法原理

YOLOv7-Tiny 作为 YOLOv7 的轻量化版本，结构精简、检测速度快。YOLOv7-Tiny 算法框架包括两个部分，骨干网络（backbone）和检测头（head），如图 1 所示，其中：k 为卷积核大小；s 为步长。网络将尺寸为 640 pixel×640 pixel 的图片作为输入层；骨干网络通过多重提取结构（ELAN-T）聚合特征信息，通过最大池化下采样（MP）进行尺度变换，以此得到多尺度下的特征信息；空间池化结构（SPP）用于加强深层特征图的感受野；检测头对特征信息多次上采样（UP），融合（Contact）浅层细节信息和深层信息，最后得到三个尺寸为 80 pixel×80 pixel、40 pixel×40 pixel 和 20 pixel×20 pixel 的有效特征层，分别用来检测小目标、中目标和大目标。解码处理得到多个预测框，输出预测边界框坐标信息、置信度和类别概率，经过非极大值抑制处理（NMS）剔除交并比（IOU）过大的预测框，保留得分最高的预测框。

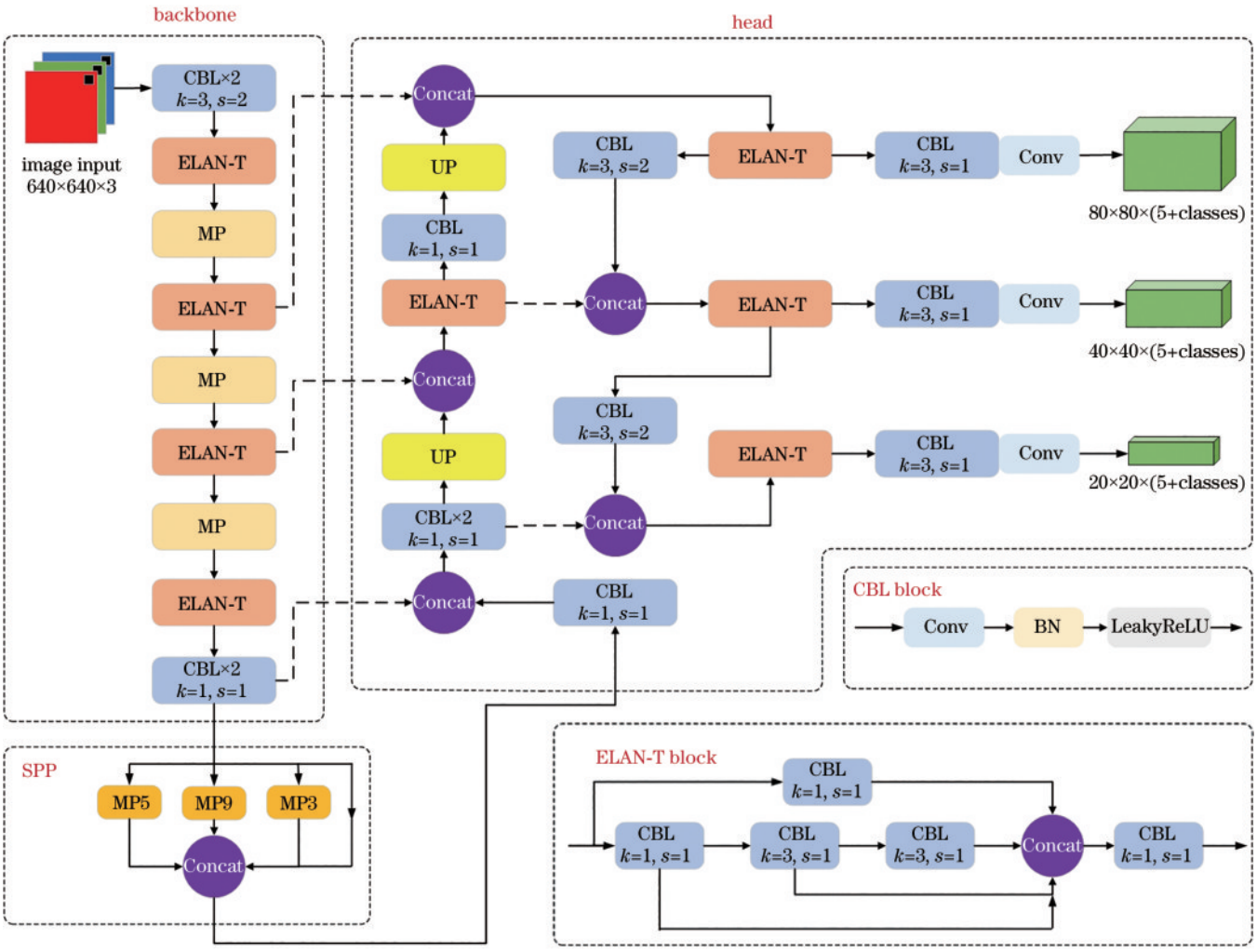
YOLOv7 系列相较于先前的 YOLOv3[19]、YOLOv4[20]，通过模型重参数化大幅提高了检测速度。模型重参数化是在推理阶段将多个计算模块合并为一个，能够极大地减少参数量和计算量，加快推理速度。如 Conv 和其后连接的批量归一（BN）层可以等价合成一个新的结构，使推理阶段在正向传播时只有一层的计算量。

2.2 融合模块设计

红外与可见光的融合关键是在提取行人热辐射特征信息并增强表现的同时，尽可能多地保留可见光背景的纹理细节信息。所提方法设计了一种在提取网络特征的同时，能够并行对同层特征进行信息融合的关联式结构，使下一层得到的特征信息不仅有上一层结构的完整输入，还包含并行提取的另外模态信息。此外，提前关注关联融合的实现也能够减少深度网络的层数，实现模型轻量化。

2.2.1 深度网络设计

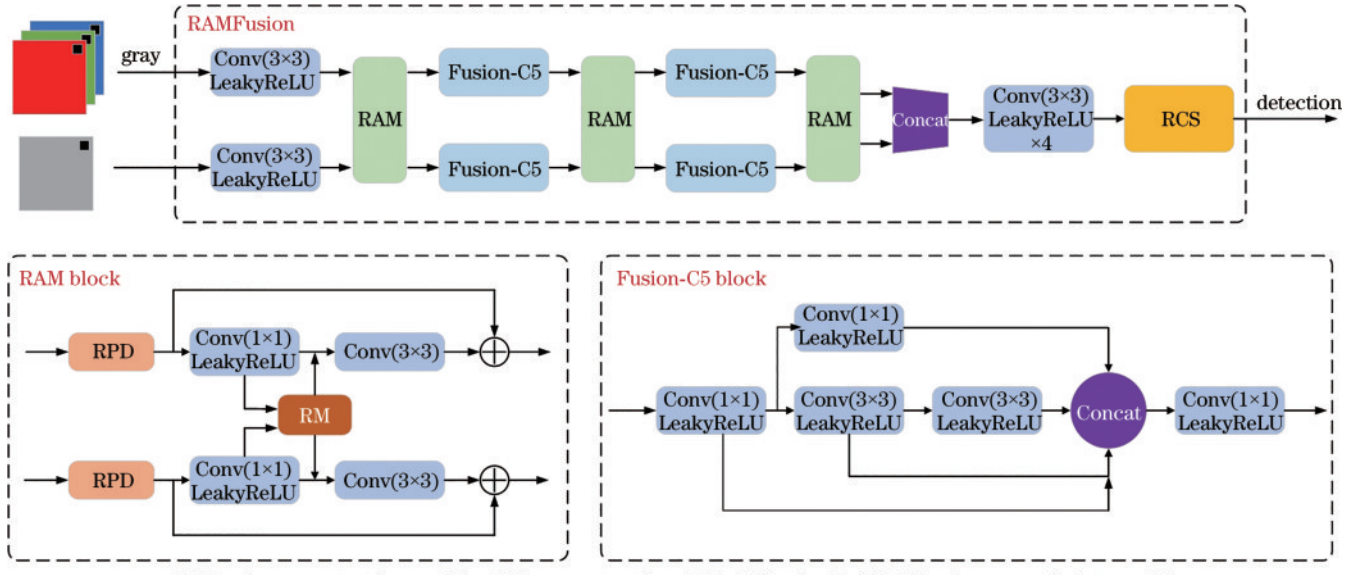
双模态融合模块如图 2 所示。首先，双模态图像分别输入双分支结构，各自的主通道使用 3×3 深度卷积升维。多重提取结构（Fusion-C5）用来加强深层特征提取；关联注意力模块（RAM）提取红外与可见光互



CBL: Conv+BN+LeakyReLU; MP: maximum pooling; SPP: spatial pyramid pooling; UP: upsampling; BN: batch normalize

图 1　YOLOv7-Tiny 网络结构图

Fig. 1　YOLOv7-Tiny network structure



RAM: relevance attention module; RCS: reconstruction; RPD: ReflectionPad2d; RM: relevance calculate module

图 2　RAM 融合网络结构图

Fig. 2　Structure diagram of RAM fusion network

补特征信息并融合，得到的关联融合特征信息通过残差结构强化表达；通道拼接（Concat）后进行多次卷积激活操作，得到融合后的灰度图；对灰度图进行图像后处理（RCS），输出的双模态融合图像作为检测网络的输入。红外与可见光的互补信息通过关联特征处理得到。

$$\boldsymbol{I}_{\mathrm{AM}}=\alpha\boldsymbol{I}_{\mathrm{ir}}+\beta\boldsymbol{I}_{\mathrm{vi}}, \tag{1}$$

式中：$\boldsymbol{I}_{\mathrm{AM}}$为关联特征；$\boldsymbol{I}_{\mathrm{ir}}$为输入的红外通道特征；$\boldsymbol{I}_{\mathrm{vi}}$为同尺度输入的可见光通道特征；$\alpha$、$\beta$为关联因子。

关联特征先进行权重分配再与单模态特征层融合，权重因子γ用来调节融合的比重。权重因子γ由两方面构成：一是关联特征自身的均值，代表对红外可见光图像共同部分的关注，增强输入图片空间转换的鲁棒性；二是突出分布，能更好地关注差异性较大的部分信息，如行人特征。计算过程为

$$\hat{\boldsymbol{I}}_x=\boldsymbol{I}_x+\min(\gamma,1)\cdot\boldsymbol{I}_{\mathrm{AM}}, \tag{2}$$

$$\gamma=\frac{1}{\max(\boldsymbol{I}'_{\mathrm{AM}})}\mathrm{Avgpool}(\boldsymbol{I}'_{\mathrm{AM}})+\frac{1}{HW}\sum\left[\boldsymbol{I}'_{\mathrm{AM}}-\mathrm{AvgPool}(\boldsymbol{I}'_{\mathrm{AM}})\right]>0, \tag{3}$$

式中：$\hat{\boldsymbol{I}}_x$为融合得到的新特征；$\boldsymbol{I}_x$为单模态特征输入；γ为融合比重的调节系数；$\boldsymbol{I}'_{\mathrm{AM}}$为对$\boldsymbol{I}_{\mathrm{AM}}$归一化后得到的结果；AvgPool(·)为全局平均池化（global average pooling）操作；H、W为输入特征的高度、宽度。

双模态图像经过一次关联注意力模块处理的效果如图 3 所示。通过对双模态输入信息[可见光图像（VI）、红外图像（IR）]加权平均处理得到关联互补信息（RM-msg）。RM-msg 优先关注目标行人和行人附近的背景纹理特征，在融合时分别给予不同的关注，新的融合图像（NEW）行人特征得到有效的边缘锐化，降低了后续检测任务的漏检率。

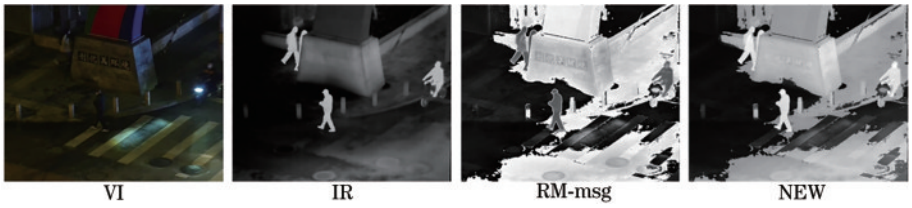


图 3　关联注意力模块的处理效果

Fig. 3　Processing effect of releted attention module

关联注意力模块整体上被设计为残差结构。一方面，残差结构能够很好地缓解网络深度增加带来的梯度消失问题，这种现象随着网络深度的增加体现得越为明显；另一方面，残差结构能有效增强深度网络性能，提高准确率。将镜像填充（RPD）得到的特征层作为残差结构的输入，后面连接卷积激活层（CL）、关联注意力模块和 3×3 Conv。关联注意力模块整体计算可描述为

$$\boldsymbol{F}_{\mathrm{C}}=\boldsymbol{R}_{\mathrm{C}}\oplus\boldsymbol{I}_x, \tag{4}$$

$$\boldsymbol{R}_{\mathrm{C}}=\mathrm{CL3}\left\langle\mathrm{RM}\left\{\mathrm{Conv1}\left[\mathrm{RPD}(\boldsymbol{I}_{\mathrm{ir}})\right],\mathrm{Conv1}\left[\mathrm{RPD}(\boldsymbol{I}_{\mathrm{vi}})\right]\right\}\right\rangle, \tag{5}$$

式中：$\boldsymbol{F}_{\mathrm{C}}$、$\boldsymbol{R}_{\mathrm{C}}$、$\boldsymbol{I}_x$分别为残差块输出特征、关联注意力提取特征、输入特征；CL3(·)、Conv1(·)分别为 3×3 卷积激活、1×1 Conv 操作；RM(·)为关联提取操作；RPD(·)为镜像填充操作；⊕表示元素相加。

2.2.2　融合损失

损失函数可以确定融合图像中保留的信息类型以及各种信息之间的比例关系。所提融合方法的损失函数包括两种损失：像素损失和梯度损失。像素损失能够约束融合图像使其和原图像的像素强度一致，保证融合前后图像的对比度不会发生过大差异。受文献[21]启发，利用光照感知子网络预先得到光照二分类处理器，再将可见光图像输入二分类网络得到的概率与融合图像进行像素损失计算。像素损失定义为

$$L_{\mathrm{pixel}}=\frac{1}{HW}\left\|P_{\mathrm{d}}\cdot(I_{\mathrm{f}}-I_1)\right\|_1+\frac{1}{HW}\left\|P_{\mathrm{n}}\cdot(I_{\mathrm{f}}-I_2)\right\|_1, \tag{6}$$

式中：P_{d}、P_{n}分别为二分类光照感知子网络得到的白天、夜晚的概率；I_{f}为融合图像；I_1、I_2分别为可见光图像、红外图像；$\|\cdot\|_1$代表 L1 损失。

同时，引入梯度损失来增强网络的约束，使融合图像能够具有更清晰的纹理特征，行人目标信息也能得到有效的边缘锐化处理。梯度损失用来计算融合图像和红外可见光输入梯度均值，表示为

$$L_{\mathrm{grad}}=\frac{1}{HW}\left\|P_{\mathrm{d}}\cdot(\nabla I_{\mathrm{f}}-\nabla I_1)\right\|_1+\frac{1}{HW}\left\|P_{\mathrm{n}}\cdot(\nabla I_{\mathrm{f}}-\nabla I_2)\right\|_1, \tag{7}$$

融合网络的总损失为

$$L=L_{\mathrm{pixel}}+\delta L_{\mathrm{grad}}, \tag{8}$$

式中：∇为 Sobel 梯度算子；δ为平衡像素损失和梯度损失的权重系数。

2.3　YOLOv7-Tiny 改进与优化

CoordConv 通过在输入特征图后增加两个坐标通道（i_x和i_y），使后续的 Conv 能够感知特征图的空间信息。通过使用 CoordConv 替换 backbone 结尾的两个 1×1 Conv 和 head 的第一个 1×1 Conv，允许行人检测网络学习平移不变性和不同程度的平移依赖性，使几

何坐标转换更加平滑。

空间分组增强注意力机制(SGE)通过解析组成特征图的每一个子特征图的重要性,可以针对性地学习和抑制噪声。重要性的提取仅仅由各个分组内全局和局部特征之间的相似性来决定,因此SGE是一个极其轻量化的神经网络注意力机制模块。SGE结构如图4所示,首先将输入的特征图在通道维度上分组(groups),通过平均池化来聚合每组的特征图空间信息,然后与每组子特征做点乘运算获取特征相关性,之后经过Sigmod函数得到归一化后的语义子特征空间注意力分数,最后与原特征图逐元素点乘获得注意力提取特征。BN可以有效避免不同样本间系数偏置造成的影响。SGE运算过程可表示为

$$\boldsymbol{F}_{\mathrm{s}}=\sigma\left\{\mathrm{BN}\left[\mathrm{AvgPool}(\boldsymbol{F})\odot\boldsymbol{F}\right]\right\}\odot\boldsymbol{F},\tag{9}$$

式中:$\boldsymbol{F}$、$\boldsymbol{F}_{\mathrm{s}}$为输入特征、SGE注意力提取特征;BN(·)为均值除以标准差的归一化操作;σ(·)为Sigmod激活函数;⊙代表点乘运算。

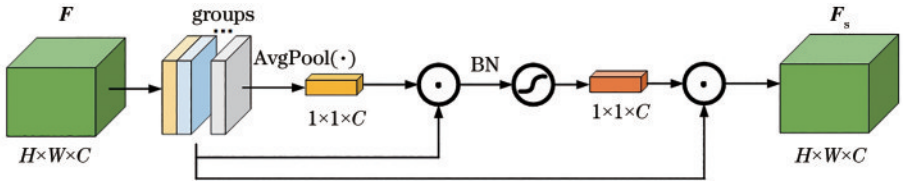


图4　SGE结构

Fig. 4　SGE structure

行人多为结群而行,多个重叠的输出检测框经常包含多个行人目标,如果使用不做改变的非极大值抑制处理(NMS)很容易将其他行人目标的检测结果全部剔除,造成行人漏检。因此,预测框处理阶段使用Soft-NMS替换NMS。

2.4　设计和优化后的网络

改进后的融合检测网络如图5所示:新网络通过RAMFusion提取和融合双模态图像的互补特征信息,突出行人特征表达;将普通1×1 Conv替换为CoordConv,引入空间感知能力;SGE模块接在ELAN-T模块之后,在增加极少参数量的情况下,增强语义区域的特征学习能力,提高检测精度;预测框处理阶段使用Soft-NMS,减少漏检。所设计的网络模型与其他融合检测网络相比,参数量更少,同时推理阶段重参数化大幅度提高了检测速度,使检测性能更加优异,符合行人实时检测标准。

3　实验结果与分析

3.1　实验设置和数据集准备

实验环境配置为:CPU i3-12100F、GPU NVIDIA GeForce RTX 3060(12 GB)、Windows 10、Python3.8、CUDA11.3、PyTorch1.7.1。

实验所用图像选自公开的LLVIP数据集[22],该数据集包含15488组红外与可见光配对图像,其中大部分都是黑暗环境下拍摄的,且所有配对图像在时间和空间上都严格校准对齐,能够很好地完成融合检测任务。考虑到原数据集对遮挡行人只标注人体部位,如骑行行人只标注头部,故对数据集部分图片重新标注,真实标注框也包含骑行车辆。从该数据集选取效果良好的4000对图像作为训练集、1000对图像作为验证集。在模型训练中,利用缩放、旋转和亮度对比度随机调整等方法对训练集数据增强处理,融合模块冻结权重,学习率采用余弦退火调整策略,初始学习率设置为0.01、最小学习率比率因子设为0.1、单次训练样本数量(batch-size)设为8、迭代次数(epoch)设为100。

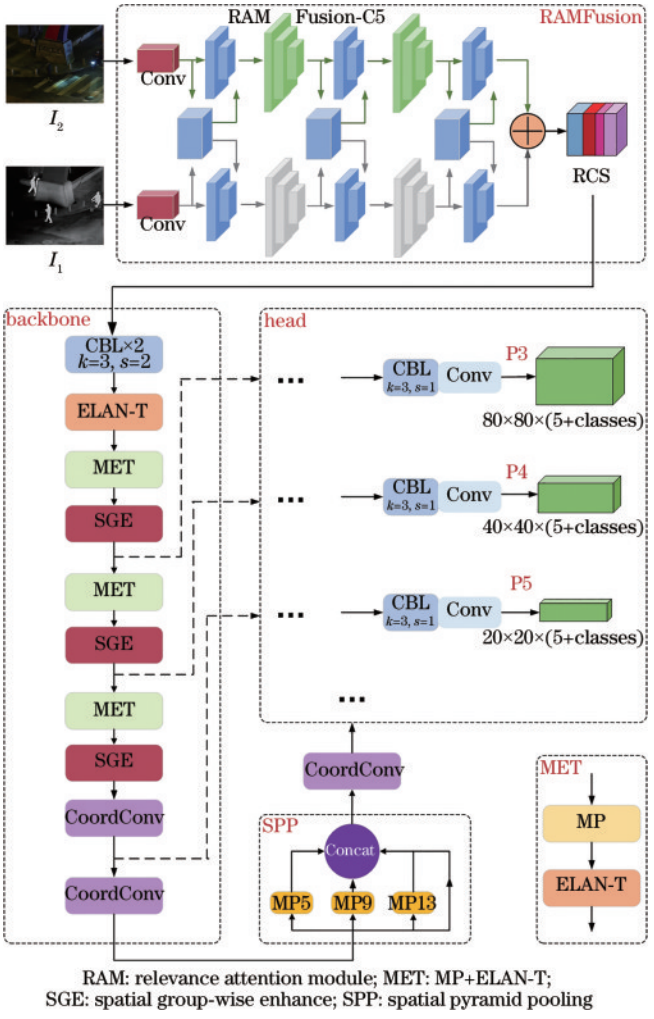


图5　设计和优化后的网络结构

Fig. 5　Designed and improved network structure

3.2 评价指标

将平均检测精度(AP)和每秒帧数(FPS)作为检测指标对所提算法进行评估,利用模型权重(weights)评估模型的部署能力。AP代表综合检测效果,考虑到边缘设备检测任务需求和数据集特点,对比IOU为0.5条件下的平均检测精度AP_{50},AP_{50}值越高综合检测效果越好。

3.3 融合方法消融实验

为验证所提融合模块对夜间行人检测的有效性,选择4种不同类型的典型融合算法与所提算法进行对比。对比算法包括:传统融合算法双树复小波变换(Wavelet)、基于深度学习的融合方法DenseFuse、NestFuse[23]和RFN-nest,对比方法均使用开源代码,检测网络使用YOLOv7-Tiny基准网络。实验结果如表1所示。

由表1可知,相较于传统融合方法Wavelet,所提方法检测精度显著提升,从83.7%提升到94.9%;对于近几年的深度网络融合方法,所提方法检测精度较DenseFuse和NestFuse分别提升了5.2个百分点和2.8个百分点,与RFN-nest仅相差0.3个百分点,但检测速度大幅提高。此外,训练迭代100次后所提融合方法召回率达到99%,即对行人基本做到了全部检测。

表1 不同融合方法的消融实验结果

Table 1 Ablation experimental results of different fusion methods

Fusion method	AP_{50} /%	Weight /MB	FPS /(frame·s^{-1})
Wavelet[12]	83.7	/	139
DenseFuse[14]	89.7	18.5	126
NestFuse[23]	92.1	24.3	98
RFN-nest[16]	95.2	33.8	53
Proposed method	**94.9**	**16.8**	**139**

3.4 改进模型消融实验

为对比YOLOv7-Tiny基准模型、YOLO系列其他模型和在YOLOv7-Tiny模型基础上改进所得模型的具体性能,设置了相关对照实验,结果如表2所示。可以看出,在基准模型YOLOv7-Tiny上,仅输入VI时,检测精度为48.1%;仅输入IR时,检测精度为76.3%。当输入为双模态图像时,嵌入设计的融合模块RAMFusion的检测精度提升到94.9%,较基准网络单独检测可见光和红外图像分别提升了46.8个百分点和18.6个百分点。当基准网络替换为同系列的YOLOv7和YOLOv7-X模型时,虽检测精度提升了1.4个百分点和1.9个百分点,但嵌入RAMFusion的YOLOv7-Tiny模型权重仅有16.8 MB,相较嵌入RAMFusion的YOLOv7和YOLOv7-X分别减小31.6 MB和65.2 MB,且检测速度快,这对于边缘设备部署和运行检测任务是非常有利的。

表2 不同模型的消融实验结果

Table 2 Ablation experimental results of different models

Method	RAM	Coord-Conv	SGE	Soft-NMS	Input	AP_{50} /%	Weight /MB	FPS /(frame·s^{-1})
YOLOv7-Tiny					VI	48.1	12.3	169
YOLOv7-Tiny					IR	76.3	12.3	169
YOLOv7-Tiny	√				VI+IR	94.9	16.8	139
YOLOv7	√				VI+IR	96.3	48.4	83
YOLOv7-X	√				VI+IR	96.8	82.0	41
Tiny-CoordConv	√	√			VI+IR	95.2	16.8	139
Tiny-SGE	√		√		VI+IR	95.8	16.9	135
Tiny-Soft-NMS	√			√	VI+IR	96.5	16.8	133
Improved model	√	√	√	√	VI+IR	**97.3**	**16.9**	**124**

在基准模型YOLOv7-Tiny上使用CoordConv替换1×1 Conv,检测精度提升0.3个百分点,说明CoordConv带来的空间感知能力对模型精度有一定提升作用;在骨干网络嵌入SGE模块,检测精度较基准网络提升0.9个百分点;引入Soft-NMS,检测精度提升1.6个百分点,这说明预测边界框由直接剔除变为降低分值再处理,对降低漏检率有着明显的帮助;同时增加SGE模块、替换CoordConv和Soft-NMS后,精度提升2.4个百分点,所作改进使模型精度有所提升,且权重基本无变化,有效实现了双模态融合检测精度和检测速度的均衡。

为更直观地比较所提算法与基准算法的检测结果,部分可视化结果如图6所示。图6(a)是可见光图像上的真实标注框(GT),图6(b)和6(c)分别是基准网络在双模态图像上的检测结果,图6(d)是基准网络单独嵌入所设计的RAM模块的双模态检测结果,图6(e)是所提算法检测结果。考虑到可视化后的效果,双模态融合检测结果均使用融合后的图像。从检测结果可以明显看出:在白天场景,能见度良好,几种算法的可见光图像检测均有较好的表现,但是红外图

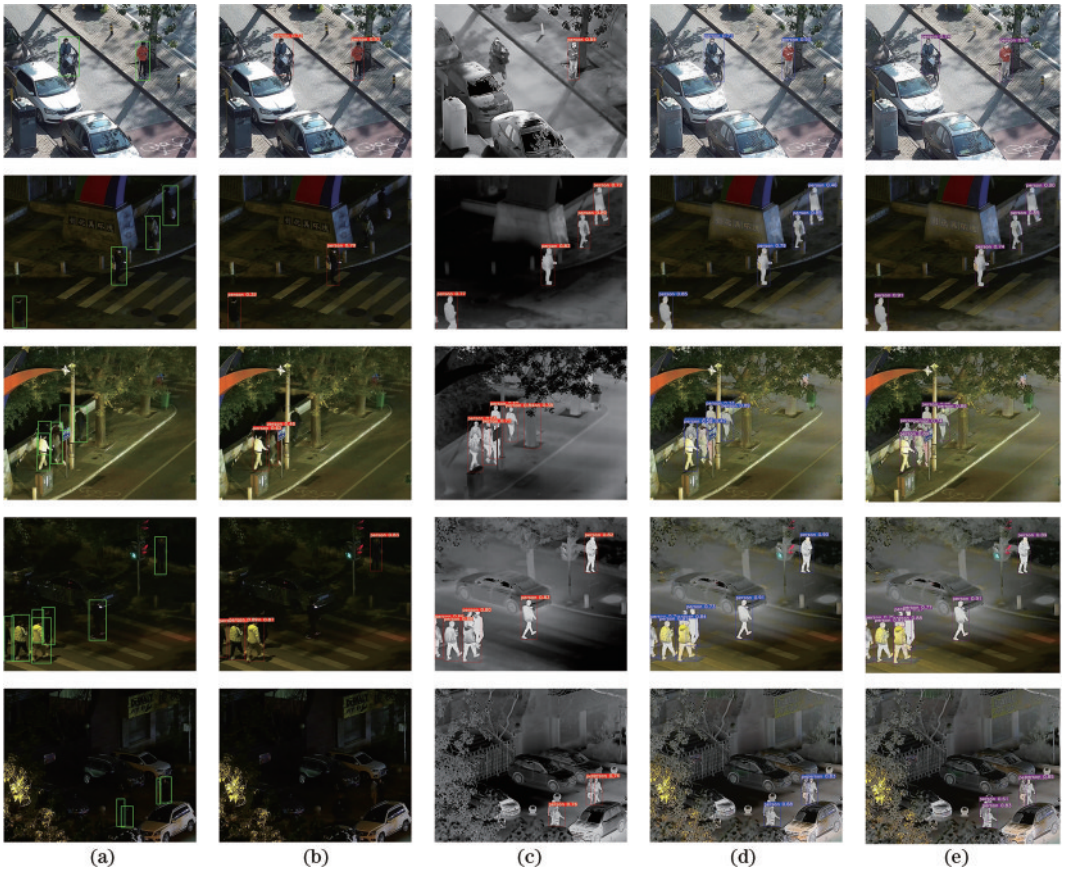


图 6 检测效果对比图。(a) VI上的GT;(b) 基准网络的VI检测结果;(c) 基准网络的IR检测结果;(d) 基准网络嵌入RAM模块后的检测结果;(e) 所提算法的检测结果

Fig. 6 Detecting results comparison. (a) GT of VI; (b) VI detection results of baseline; (c) IR detection results of baseline; (d) detection results of baseline after the RAM module is embedded; (e) detection results of the proposed algorithm

像目标不明显;当能见度降低时,基准算法在可见光图像检测时效果下降,出现检测错误、漏检;增加RAM模块后,通过融合双模态特征,红外的热辐射信息与可见光信息有效互补,能检测完全黑暗环境的行人目标;在行人较多且有遮挡重叠时,其他算法检测时有行人目标未被检测到,而所提改进算法能够有效降低漏检率,准确检测到每个行人目标。

4 结 论

提出了一种平衡双模态融合检测精度和检测速度的双模态行人检测模型。该模型通过设计的RAMFusion提取和聚合红外与可见光互补信息,通过改进YOLOv7-Tiny检测网络提高检测精度,不仅解决了单模态检测器在暗光场景下无法检测到目标的问题,还增强了双模态行人特征信息,使行人检测性能显著提高。实验结果表明,所提模型权重占用为16.9 MB、速度达到124 frame/s、检测精度达到97.3%,能更好地部署到边缘设备,应用于行人实时检测任务。研究结果为今后夜间等低能见度场景的检测任务提供了参考。

参 考 文 献

[1] 罗艳, 张重阳, 田永鸿, 等. 深度学习行人检测方法综述[J]. 中国图象图形学报, 2022, 27(7): 2094-2111.
Luo Y, Zhang C Y, Tian Y H, et al. An overview of deep learning based pedestrian detection algorithms[J]. Journal of Image and Graphics, 2022, 27(7): 2094-2111.

[2] Xu F L, Liu X, Fujimura K. Pedestrian detection and tracking with night vision[J]. IEEE Transactions on Intelligent Transportation Systems, 2005, 6(1): 63-71.

[3] Ge J F, Luo Y P, Tei G. Real-time pedestrian detection and tracking at nighttime for driver-assistance systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(2): 283-298.

[4] Ren S Q, He K M, 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.

[5] He K M, Gkioxari G, Dollar P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[6] 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.

[7] 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.

[8] 邵延华, 张铎, 楚红雨, 等. 基于深度学习的YOLO目标检测综述[J]. 电子与信息学报, 2022, 44(10): 3697-3708.
Shao Y H, Zhang D, Chu H Y, et al. A review of YOLO object detection based on deep learning[J]. Journal of Electronics & Information Technology, 2022, 44(10): 3697-3708.

[9] 何自芬, 陈光晨, 陈俊松, 等. 多尺度特征融合轻量化夜间红外行人实时检测[J]. 中国激光, 2022, 49(17): 1709002.
He Z F, Chen G C, Chen J S, et al. Multi-scale feature fusion lightweight real-time infrared pedestrian detection at night[J]. Chinese Journal of Lasers, 2022, 49(17): 1709002.

[10] 孙颖, 侯志强, 杨晨, 等. 基于双模态融合网络的目标检测算法[J]. 光子学报, 2023, 52(1): 0110002.
Sun Y, Hou Z Q, Yang C, et al. Object detection algorithm based on dual-modal fusion network[J]. Acta Photonica Sinica, 2023, 52(1): 0110002.

[11] 刘子龙, 沈祥飞. 融合Lite-HRNet的Yolo v5双模态自动驾驶小目标检测方法[J]. 汽车工程, 2022, 44(10): 1511-1520, 1536.
Liu Z L, Shen X F. Yolo v5 dual-mode automatic driving small target detection method combining lite-HRNet[J]. Automotive Engineering, 2022, 44(10): 1511-1520, 1536.

[12] Ben Hamza A, He Y, Krim H, et al. A multiscale approach to pixel-level image fusion[J]. Integrated Computer-Aided Engineering, 2005, 12(2): 135-146.

[13] Wang L, Li B, Tian L F. EGGDD: an explicit dependency model for multi-modal medical image fusion in shift-invariant shearlet transform domain[J]. Information Fusion, 2014, 19: 29-37.

[14] Li H, Wu X J. DenseFuse: a fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.

[15] Xu H, Ma J Y, Jiang J J, et al. U2Fusion: a unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(1): 502-518.

[16] Li H, Wu X J, Kittler J. RFN-Nest: an end-to-end residual fusion network for infrared and visible images[J]. Information Fusion, 2021, 73: 72-86.

[17] 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[EB/OL]. (2022-07-06)[2023-02-03]. https://arxiv.org/abs/2207.02696.

[18] Liu R, Lehman J, Molino P, et al. An intriguing failing of convolutional neural networks and the CoordConv solution[EB/OL]. (2018-07-09)[2023-02-03]. https://arxiv.org/abs/1807.03247.

[19] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2023-02-03]. https://arxiv.org/abs/1804.02767.

[20] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23)[2023-02-03]. https://arxiv.org/abs/2004.10934.

[21] Tang L F, Yuan J T, Zhang H, et al. PIAFusion: a progressive infrared and visible image fusion network based on illumination aware[J]. Information Fusion, 2022, 83/84: 79-92.

[22] Jia X Y, Zhu C, Li M Z, et al. LLVIP: a visible-infrared paired dataset for low-light vision[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), October 11-17, 2021, Montreal, BC, Canada. New York: IEEE Press, 2021: 3489-3497.

[23] Li H, Wu X J, Durrani T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models[J]. IEEE Transactions on Instrumentation and Measurement, 2020, 69(12): 9645-9656.