基于改进YOLOv5的双模态图像融合行人检测方法

王维锋[1],邵琳骞[1],黄建鑫[1],李洪囤[2],张日民[3]

(1.河海大学 土木与交通学院,南京 210098;2.交通运输部科学研究院,北京 100029;3.苏交科集团 智慧交通研究所,南京 210019)

摘 要:为解决低照度和局部遮挡场景下行人检测漏检率高、红外与可见光双模态融合图像检测速度慢的问题,提出了一种双模态融合图像的轻量化行人检测方法。利用基于语义感知的实时图像融合网络对红外与可见光图像进行特征级融合,选取YOLOv5为基准模型,采用GhostNet优化其主干网络,在空间金字塔池化模块输出位置添加全局注意力机制,将损失函数由CIoU调整为EIoU以提高收敛速度与精度。采用开放数据集LLVIP开展实验研究,实验结果表明,相较于改进前的YOLOv5模型,本文方法在模型权重减少32.1%的情况下,每秒图像处理帧数增加4帧,平均检测精度达到94.7%,具有较快的检测速度和良好的检测精度。

关键词:智能交通;行人检测;轻量化;注意力机制;双模态;YOLOv5

中图分类号:U491.5 **文献标志码**:A **文章编号**:1671-5497(2026)06-1727-10

DOI:10.13229/j.cnki.jdxbgxb.20241288

Pedestrian detection via dual-modal image fusion based on improved YOLOv5

WANG Wei-feng[1],SHAO Lin-qian[1],HUANG Jian-xin[1],LI Hong-dun[2],ZHANG Ri-min[3]

(1. *College of Civil and Transportation Engineering, Hohai University, Nanjing* 210098, *China*; 2. *China Academy of Transportation Sciences, Beijing* 100029, *China*; 3. *Intelligent Transportation Research Institute, JSTI GROUP, Nanjing* 210019, *China*)

Abstract: Adopting single modal to detect pedestrians has a high missed detection rate under the circumstance of poor illumination and partial occlusion, while dual-modal has low real-time detection speed. To address this issue, this study proposed a lightweight method using dual-modal to detect pedestrians based on an improved deep learning model. Firstly, a Semantic-Aware Real-Time Image Fusion Network (SeAFusion) based on semantic perception was employed to fuse infrared and visible-light images at the feature level. Secondly, a lightweight network, GhostNet, was integrated into the backbone network of the YOLOv5 baseline model. Thirdly, following the Spatial Pyramid Pooling Fast (SPPF)

收稿日期:2024-12-02.

基金项目:中央高校基本科研业务费专项资金项目(B240201168),江苏省交通运输科技项目(2024Y19-2).

作者简介:王维锋(1979-),男,教授,博士.研究方向:交通仿真,人工智能感知,交通大数据.
E-mail: wangweifeng@hhu.edu.cn

通信作者:李洪囤(1973-),男,副研究员,硕士.研究方向:交通运输大数据,交通信息化.E-mail: lihd@motcats.ac.cn

module in the backbone network, a Global Attention Mechanism (GAM) was added to adaptively acquire global multi-dimensional features and further reduce information diffusion. To improve the convergence speed and accuracy of model solving, the Complete Intersection Over Union (CIoU) loss function was replaced by the Efficient Intersection Over Union (EIoU). Finally, this study employed the open dataset LLVIP for experimental validation. The results show that compared with the unimproved YOLOv5 model, the proposed method can increase the image processing rate by 4 frames per second and achieve a mean average precision of 94.7% for pedestrian detection with a 32.1% reduction in model weights.

Key words: intelligent transportation; pedestrian detection; lightweight; attention mechanism; dual-modal; YOLOv5

0 引 言

行人检测是计算机视觉技术在交通目标识别领域的重要研究方向,然而,可见光单模态图像在低照度条件下无法区分行人与环境的外观特征[1,2],红外单模态图像存在作用距离短、分辨率低、纹理细节较少等问题[3,4],基于可见光与红外双模态图像的行人检测方法也面临网络结构复杂和实时检测效率低的挑战,因此,对行人检测算法及结构进行优化、提高检测精度的同时保证实时性仍是未来研究的重点[5]。

传统的行人检测方法主要根据行人与背景特征进行识别,使用滑动窗口技术生成可能包含行人的候选框,并采用SIFT、HOG、Haar等方法从候选框中提取行人外观特征,最后通过Adaboost、SVM等分类器检测行人[6,7],但是存在鲁棒性差、检测效果不佳等问题[8]。针对上述问题,相关学者提出了基于深度学习的目标检测算法,其主要分为二阶段模型(如Faster R-CNN[9]、Mask R-CNN[10])和一阶段模型(如SSD[11]、YOLO[12,13])。研究人员基于二阶段模型进行了一系列可见光与红外双模态图像行人检测研究工作[14,15],然而二阶段模型的网络结构过深、模型参数量大、推理速度有限,无法满足实时检测要求。现阶段的双模态图像行人检测研究工作中主要采用一阶段模型,Xue等[16]在YOLOv3基础上构建双模态加权模块进行夜间行人识别,有效提升了夜间场景下行人的检测精度;Li等[17]在YOLOv5基础上采用Ghost卷积替代普通卷积进行行人识别,该方法有效减少模型参数量,但存在泛化效果不佳的问题;孙文财等[18]通过Ghost瓶颈模块实现了算法复杂度与融合质量之间的平衡,可在复杂天气条件下显著提升道路目标检测指标性能;刘子龙等[19]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,在检测道路中的小目标和密集目标方面表现优异,然而,采用的像素级融合方法未能有效整合丰富的红外与可见光特征信息,导致行人检测效果不佳。Falaschetti等[20]采用基于Tucker张量分解的卷积神经网络压缩技术优化YOLOv3-tiny,王宏志等[21]通过增加小目标检测层和解耦检测头对YOLOv5算法进行改进,为实时检测提供了重要参考。以上多模态融合检测方法虽提取了可见光与红外双模态图像行人特征信息,但网络结构过深、模型参数量大、推理速度有限,嵌入的目标检测网络无法满足实时检测要求。

为保证双模态图像的高质量融合,本文使用级联视觉任务的图像融合方法SeAFusion[22]对红外-可见光图像进行特征级融合;由于YOLOv5模型主要用于可见光图像检测任务,因此需要基于双模态融合图像对模型进行适应性改进。本文使用轻量化网络GhostNet[23]替换YOLOv5原有的主干网络,降低其网络参数量;为平衡引入轻量化网络造成的信息弥散,融合GAM[24]注意力机制以细化特征图,放大全局跨维度的相互作用,提高检测任务的精度;最后,通过将损失函数改进为EIoU,解决YOLOv5原有损失函数CIoU因长宽比例损失而无法解耦的问题,进而加快双模态图像行人检测模型在训练过程中的收敛速度。

1 行人检测总体框架

针对低照度下行人目标检测效果不佳、双模态融合图像检测方法存在参数量过大、导致实时检测能力较弱等问题,本文采用SeAFusion进行图像融合,提出了改进YOLOv5的双模态图像行人检测框架,如图1所示。

该框架包含双模态融合图像算法和改进的 YOLOv5 算法，其中融合端采用梯度残差密集块（GRDB）获取双模态融合图像；检测端采用改进的 YOLOv5 算法，其由 Backbone 网络、Neck 网络、Head 网络三部分构成。改进的 YOLOv5 算法中采用轻量化网络 GhostNet 替换主干网络中的普通卷积，降低模型的复杂度；在空间特征金字塔池化模块（SPPF）中加入 GAM 注意力机制，提取全局多维度信息；损失函数采用 EIoU 代替传统 CIoU，加快模型收敛速度。

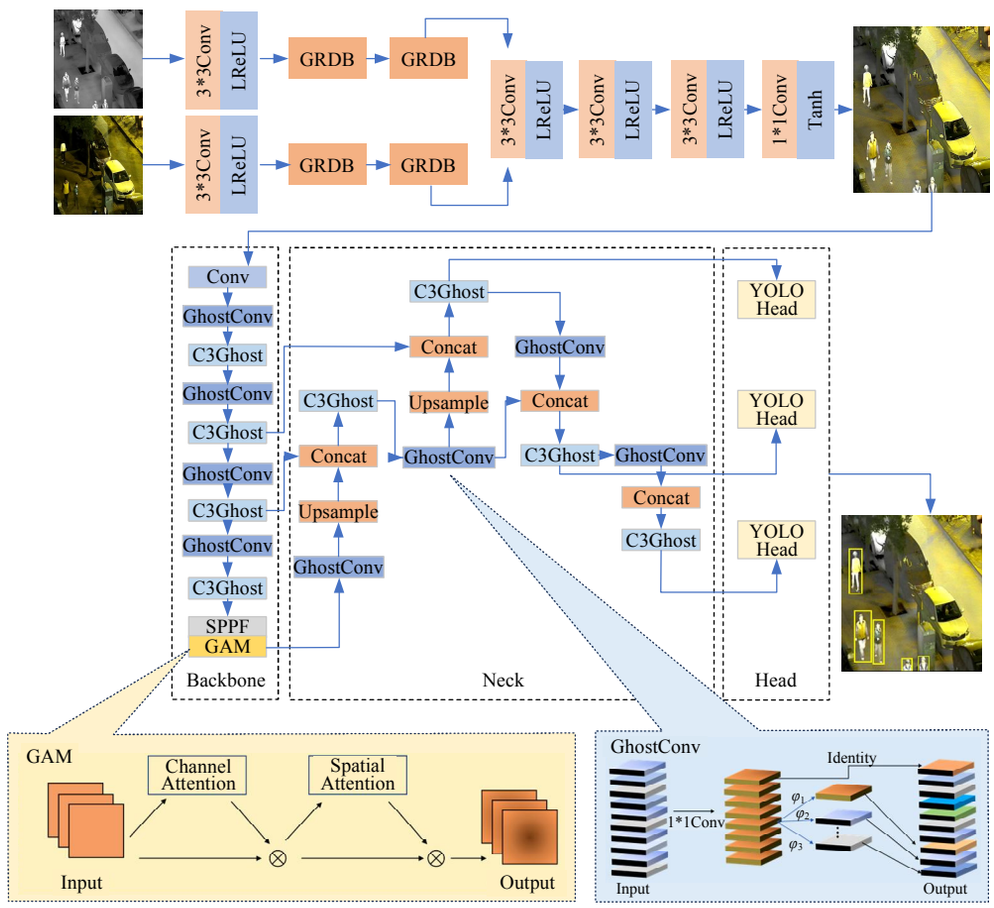


图 1　基于改进 YOLOv5 的双模态融合图像行人检测框架

Fig. 1　Frame workof pedestrian detection using a dual-modal fusion image based on improved YOLOv5

2　行人检测模型构建

2.1　双模态融合模型

为融合可见光图像与红外图像各自具有的丰富特征进而实现行人检测任务，本文采用 SeAFusion 图像实时融合网络对行人红外-可见光数据进行特征级融合。现有的融合算法大多关注于上游融合图的视觉质量和评价指标，而忽略如目标检测、语义分割等下游高级视觉任务，SeAFusion 算法弥合图像融合与高级视觉任务的差距，通过将图像融合模块和语义分割模块进行级联，利用语义损失的反向传播指导图像融合网络训练，有效提升融合图像在高层视觉任务中的性能，并通过设计 GRDB 增强融合网络提取纹理特征的能力。SeAFusion 网络结构如图 2 所示。

双模态图像融合的具体方法如下：

（1）输入端：给定配准的红外-可见光图像 I_{ir}、I_{vi}。

（2）图像融合模块：利用嵌入 GRDB 的特征提取模块进行轻量化双模态特征级融合，首先使用特征提取模块 E_F 从红外与可见光图像中提取深层特征，其表示为：

$$\{F_{ir}, F_{vi}\}=\{E_F(I_{ir}), E_F(I_{vi})\} \tag{1}$$

式中：F_{ir} 和 F_{vi} 分别为红外特征与可见光特征。

此外，E_F 中嵌入的 GRDB 结构可以增强对细颗粒度特征的提取能力，给定 GRDB 的输入特征为 F^i，其输出特征 F^{i+1} 表示为：

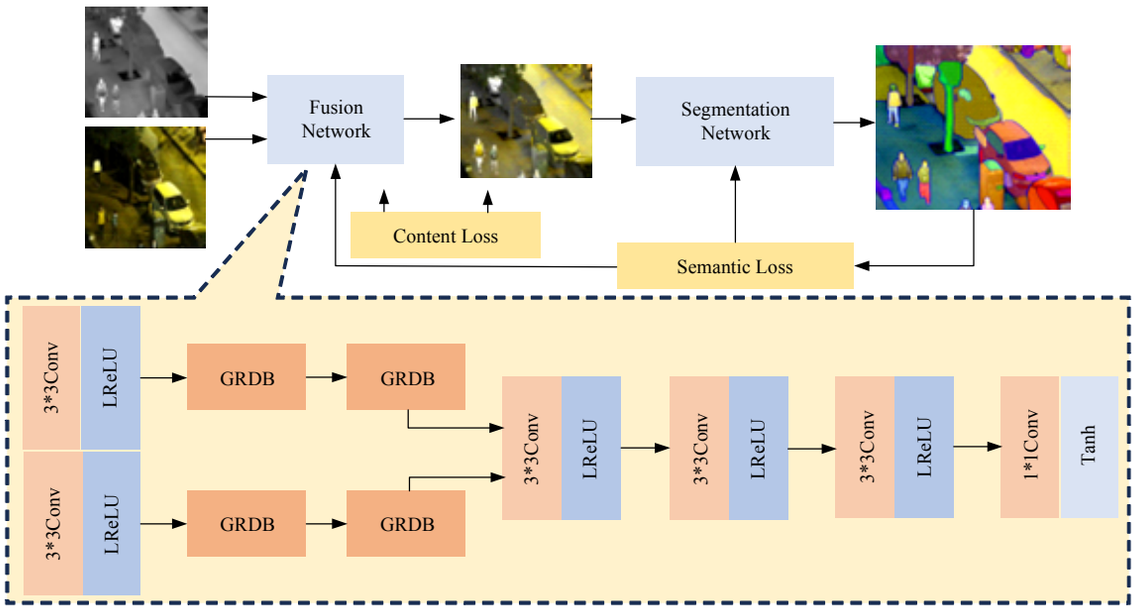


图 2 SeAFusion 网络结构

Fig. 2 Network architecture of SeAFusion

$$F^{i+1}=GRDB(F^i)=Conv^n(F^i)\oplus Conv(\nabla F^i) \tag{2}$$

式中:$Conv$ 表示卷积层;$Conv^n$ 表示 n 级联卷积层;$\oplus$ 表示元素求和。

其次,使用特征融合与图像重建模块对融合后的图像进行重建,其中特征融合模块与图像重建模块分别表示为:

$$F_f=C(F_{ir},F_{vi}) \tag{3}$$

$$I_f=R_I(F_f) \tag{4}$$

式中 C 为通道拼接;F_f 为融合特征;R_I 为图像重建模块;I_f 为最终融合图像。

融合结果使用内容损失 $L_{content}$ 作为量化指标,其由强度损失 L_{int} 和纹理损失 $L_{texture}$ 两部分组成。内容损失表示为:

$$L_{content}=L_{int}+L_{texture} \tag{5}$$

(3)语义分割模块:使用分割模型 N_s 对融合图像 I_f 执行分割,分割结果的 I_s 与语义标签 L_s 之间的损失 $L_{semantic}$ 可反映语义信息的丰富程度。语义分割过程表示为:

$$I_s=N_s(I_f) \tag{6}$$

语义损失 $L_{semantic}$ 表示为:

$$L_{semantic}=\varepsilon(I_s,L_s) \tag{7}$$

式中:ε 为误差函数。

(4)联合损失函数:联合损失函数 L_{joint} 可通过反向传播指导图像融合模型训练,L_{joint} 表示为:

$$L_{joint}=L_{content}+\beta L_{semantic} \tag{8}$$

式中:β 为常数,其取值随模型训练自适应变化。

2.2 改进的 YOLOv5 模型

2.2.1 轻量化网络模型

GhostNet 是一种轻量级神经网络结构,通过计算量更小的操作代替普通卷积生成常规特征提取方法中的冗余特征图。在模型网络的特征层中充足或者冗余的信息总是可以保证对输入数据的全面理解,在行人图像的特征层之间有众多信息是相似的,这些相似的特征层能够全面表达输入数据,因此不必为冗余的相似信息提供过多的计算资源,本文使用由 Ghost 卷积和 GhostBottleneck 构成的 C3Ghost 替代了 YOLOv5 模型中 Backbone 的 C3 模块,以降低计算成本。Ghost 模块的思想是将普通卷积分为两步:第 1 步,对输入特征图使用 1×1 普通卷积进行特征浓缩;第 2 步,使用 1×1 深度卷积对每个通道特征图提取特征,获得另一半特征图,最后将两次卷积的输出特征图在通道维度上堆叠,得到与传统卷积层同等特征表达能力的相似特征图。假设输入特征图形状为 $[H,W,C]$,H 和 W 分别为特征图高度和宽度,px;C 为输入通道数,Ghost 卷积参数量为:

$$G_{params}=G_{C_{params}}+G_{GC_{params}} \tag{9}$$

$$G_{C_{params}}=K\times K\times C\times(\frac{C}{2}) \tag{10}$$

$$G_{GC_{params}}=K\times K\times(\frac{C}{2}) \tag{11}$$

式中:G_{params} 为 Ghost 卷积参数量;$G_{C_{params}}$ 为 Ghost

中普通卷积参数量；$G_{GC_{params}}$为Ghost中深度卷积的参数量；K为卷积核的尺寸。

传统卷积参数量为：

$$C_{params}=K\times K\times C\times C \tag{12}$$

式中：C_{params}为传统卷积的参数量。

用式(9)除以式(12)，得到：

$$\frac{G_{params}}{C_{params}}=\frac{1}{2}+\frac{1}{2C} \tag{13}$$

由于YOLOv5网络中通道数C的取值一般在64以上[19]，因此根据式(13)可以推出，Ghost卷积的参数量约为传统卷积的1/2，利用Ghost模块替换YOLOv5原有主干网络中的普通卷积可以有效减少模型参数量。

2.2.2　融合GAM注意力机制

传统注意力方法，如编码器-解码器残余注意模块和压缩-激励网络等，可用于细化特征图以获得较好的性能。然而，这些方法在抑制非重要的像素方面效率较低，新出现的自适应卷积注意力模块和时域自适应模块虽然跨越空间和通道维度，但未考虑通道与空间的交互作用。本文使用融合全局注意力机制GAM方法，通过综合考虑空间和通道维度，既实现在多个维度上捕捉重要特征，也同时放大全局跨维度的相互作用。全局注意力机制在行人目标检测任务中，通过对整个图像信息进行加权处理，从而在复杂光照环境下将行人与背景进行区分，实现更为准确的行人识别任务。

本文采用图3所示的GAM结构，其通过输入特征图F_1，借助通道注意力机制模块对输入特征图进行校正，得到中特征图F_2，然后经过空间注意力机制模块再次校正，得到输出特征图F_3。

中特征图F_2和输出特征图F_3分别表示为：

$$F_2=M_c(F_1)\otimes F_1 \tag{14}$$

$$F_3=M_s(F_2)\otimes F_2 \tag{15}$$

式中：M_c和M_s分别为通道注意力和空间注意力特征图谱；$\otimes$表示元素求积。

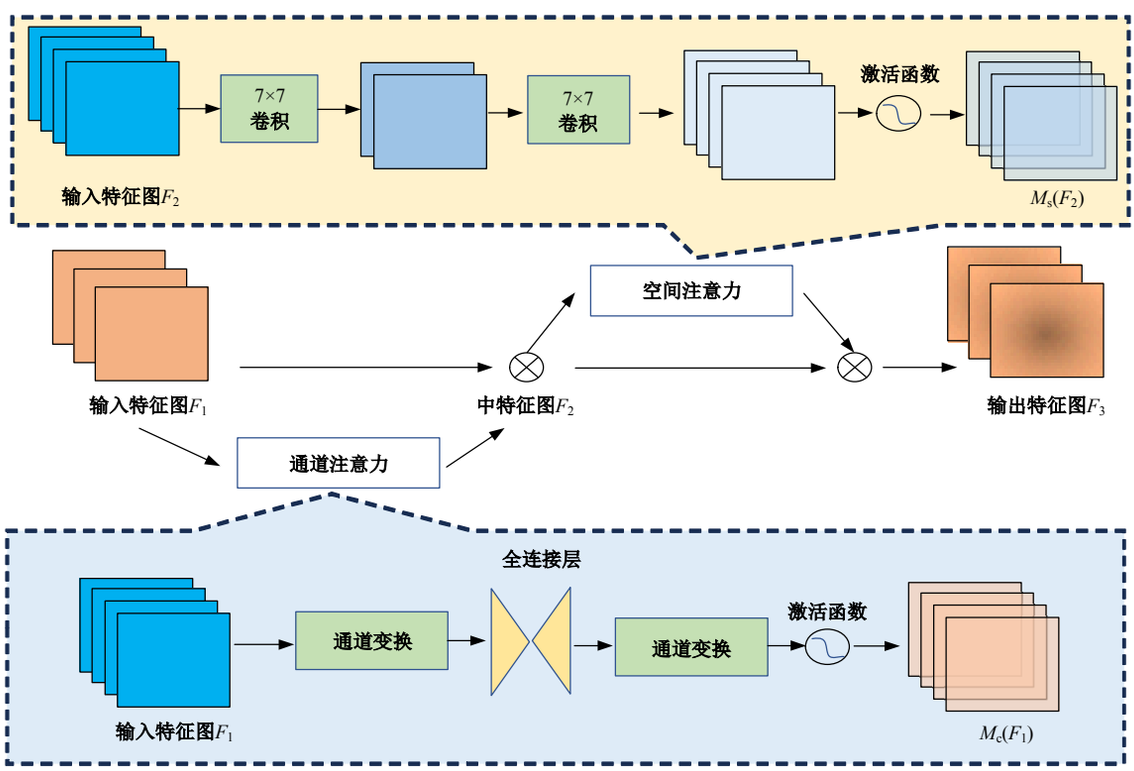


图3　GAM结构

Fig. 3　GAM structure

2.2.3　EIoU损失函数

在改进YOLOv5模型中采用EIoU损失函数替换传统的CIoU损失函数。真实边界框为测试集中已标注的对应行人边界框，预测边界框为目标检测模型预测推理出的行人边界框，EIoU损失函数结合预测边界框和真实边界框之间的交并比IoU和欧氏距离，相比CIoU损失函数，其具有更强的梯度鲁棒性。特别在双模态行人图像拥有更多特征信息的情况下，EIoU损失函数可以更好的处理预测边界框中的小误差，从而提高行人检测

准确性,有助于模型更快收敛并取得更好的性能。

CIoU损失计算如下:

$$L_{\mathrm{CIoU}}=1-\mathrm{IoU}+\frac{\rho^2(b,b^{gt})}{C^2}+\alpha v \tag{16}$$

式中:IoU为预测框与真实框的交并比;$\rho^2(b,b^{gt})$为预测框与真实框的欧式距离;α为权重系数;v为度量长宽比的相似性;C为覆盖预测框与真实框的最小外接矩形对角线长度。

EIoU损失计算如下:

$$L_{\mathrm{EIoU}}=L_{\mathrm{IoU}}+L_{\mathrm{dis}}+L_{\mathrm{asp}}=1-IoU+\frac{\rho^2(b,b^{gt})}{C^2}+\frac{\rho^2(w,w^{gt})}{C_{\mathrm{w}}^2}+\frac{\rho^2(h,h^{gt})}{C_{\mathrm{h}}^2} \tag{17}$$

式中:C_{w}、C_{h}分别为同时包含预测框与真实框的最小外接矩形的宽度和高度;$\rho^2(w,w^{gt})$与$\rho^2(h,h^{gt})$分别为预测框与真实框之间的宽度差值与高度差值。

3 实验与结果分析

3.1 数据集和实验环境

本文选择公开数据集LLVIP[25]用于模型训练与测试,LLVIP数据集是在时间和空间上已完成配准的15 488对红外与可见光图像,其分辨率为1 080像素×720像素,行人在图像中的位置和类别信息由Labelimg工具标注得到。采用SeAFusion算法将配准的红外与可见光图像进行特征级融合,得到具有可见光纹理特征与红外边缘特征的融合图像,不同时段双模态融合示例如图4所示。



图4 不同时段双模态融合示例

Fig. 4 Example of dual-modal image fusion at different time periods

本文模型训练过程中使用的显卡为GeForce RTX 2080 ti(22 GB),CPU型号为Intel(R) Core(TM) i5-12490F,所使用深度学习框架为PyTorch 1.11.0+CUDA 11.5;模型在GPU上训练时使用SGD为优化器,设置初始学习率为0.01,循环学习率为0.2,衰减为0.005,Batch_size为32,训练轮次为150次,为提高模型训练效率,输入图片均重置为统一尺寸(640像素×640像素)。

3.2 评价指标

本文采取的模型评价指标分别为精确率、召回率、平均精度、模型权重大小,具体定义如下。

精确率$P_{\mathrm{Precision}}$为检测出的正样本数占所有检测出的样本数的比例:

$$P_{\mathrm{Presision}}=\frac{T_{\mathrm{p}}}{T_{\mathrm{p}}+F_{\mathrm{p}}} \tag{18}$$

式中:T_{p}为样本中原是正例且被正确划分为正例的数量;F_{p}为样本中原是负例但被错误划分为正例的数量。

召回率R_{recall}为检测出的正样本数占所有实际正样本数的比例:

$$P_{\mathrm{recall}}=\frac{T_{\mathrm{p}}}{T_{\mathrm{p}}+F_{\mathrm{n}}} \tag{19}$$

式中:F_{n}为样本中原是正例但被错误划分为负例的数量。

平均精度R_{mAP}为行人检测的平均精度:

$$R_{\mathrm{mAP}}=\frac{1}{N}\sum_{i=1}^{N}AP_i \tag{20}$$

$$AP=\int_0^1 P(r)\mathrm{d}r \tag{21}$$

式中:N为类别数量;AP为精确率-召回率曲线与坐标轴围成的面积;AP_i为第i个类别的平均精度。当IoU达到50%时,模型计算的平均精度记为mAP50。

模型权重大小W_{weight}计算方法如下:

$$W_{\mathrm{weight}}=\frac{P_{\mathrm{params}}\times m}{1\,024^2} \tag{22}$$

式中:P_{params}为模型的参数量;m为每个参数的字节数,byte。

3.3 模型训练

按照8∶2的比例将数据集划分为训练集(12 390张融合图像)和测试集(3 098张融合图像)。

本实验旨在比较改进前的YOLOv5模型和本文改进后的模型在多模态融合图像行人检测任务中的表现,并分析改进模型的优化效果。通过上述图像融合后的LLVIP数据集计算mAP50指标,并比较两种模型的表现。IoU损失对比如图5所示,从损失函数变化曲线可知,本文改进模型所使用的EIoU在前20个训练轮次的损失下降速度更快且曲线更平稳,YOLOv5的损失曲线下降相对较慢且波动性较大。经过110个轮次的训练,

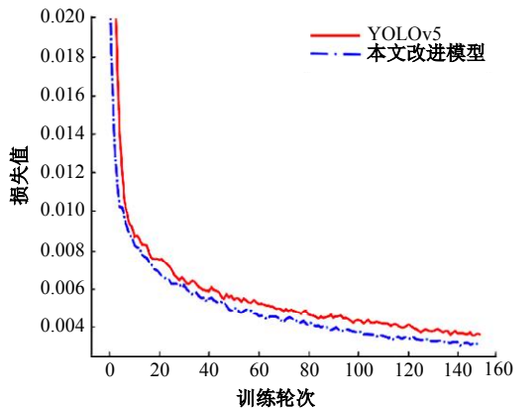


图 5　损失函数对比

Fig. 5　Comparison of loss functions

本文改进模型的损失函数值下降至0.003 5,而YOLOv5的损失函数值下降至0.004 1,说明本文改进模型使用的EIoU相比于YOLOv5的CIoU具有更优的收敛效果。

将测试集输入训练好的模型中对行人进行检测,并采用Grad CAM绘制YOLOv5模型改进前、后的热力图,如图6所示。对比发现,改进前的模型对于行人目标的关注范围较小,甚至存在关注区域缺失问题;改进后的模型能够更为精准地关注目标位置,且对于行人身体存在部分重叠的图像,也能有效提取更完整的行人轮廓特征。

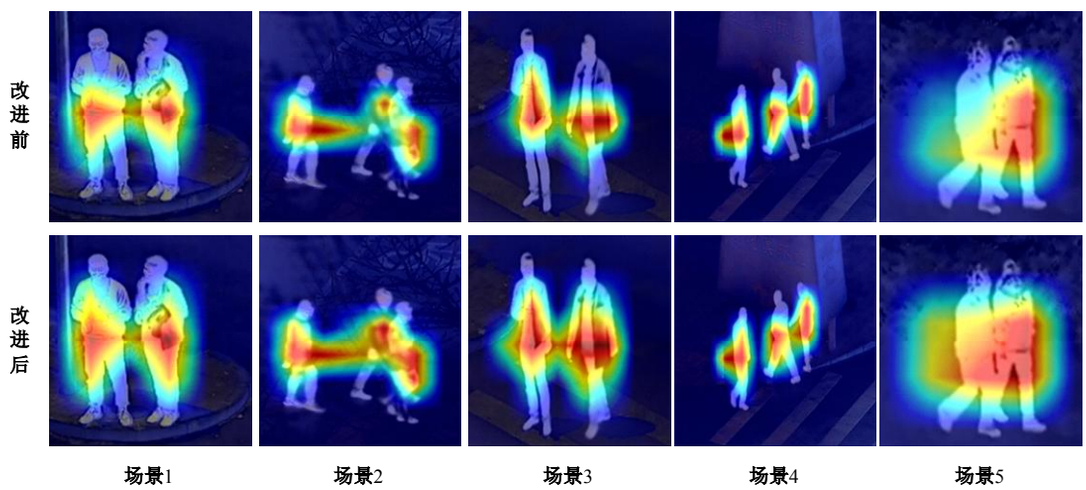


图 6　模型改进前、后的行人检测热力图

Fig. 6　Heatmap of pedestrian detection before and after model improvement

3.4　实验结果与分析

为验证图像融合后的检测效果,选取单模态可见光图像、单模态红外图像、双模态融合图像进行检测结果对比,见表1。

表 1　图像融合前后检测结果对比

Table 1　Comparison of detection results before and after image fusion

模型	输入数据	精确率/%	召回率/%	mAP50/%	模型权重大小/MB
YOLOv5	单模态可见光图像	86.0	82.5	88.2	14.0
YOLOv5	单模态红外图像	94.9	86.6	94.4	14.0
本文改进模型	双模态融合图像	95.7	86.9	94.7	9.5

图像融合前后的检测效果对比如图7所示,图7(a)~(c)分别为可见光图像检测效果、可见光图像配对的红外图像检测效果和双模态融合图像经本文改进模型输出的检测效果。

图7第一行的结果显示,在夜间光照昏暗的情况下,可见光图像存在行人目标漏检情况,而红外图像、双模态融合图像均能检测出行人目标,且二者检测结果的置信度相同;图7第二行的结果显示,在夜间灯光昏暗且有路侧树木遮挡的情况下,可见光图像仍然存在行人目标漏检现象,双模态融合图像对行人被局部遮挡的检测置信度为0.76,相比红外图像检测提升7%;图7第三行的结果显示,在夜间灯光极度昏暗且有路侧树木遮挡的情况下,可见光未检测出行人目标,双模态融合图像对行人被严重遮挡的检测置信度为0.73,相比红外图像检测提升14%。因此,双模态图像检测方法在行人目标检测方面具有更强的鲁棒性。

表2为本文改进模型与常用模型的检测结果对比,表2中输入数据均为双模态融合图像。由表2可知,本文改进模型用于行人检测的平均精

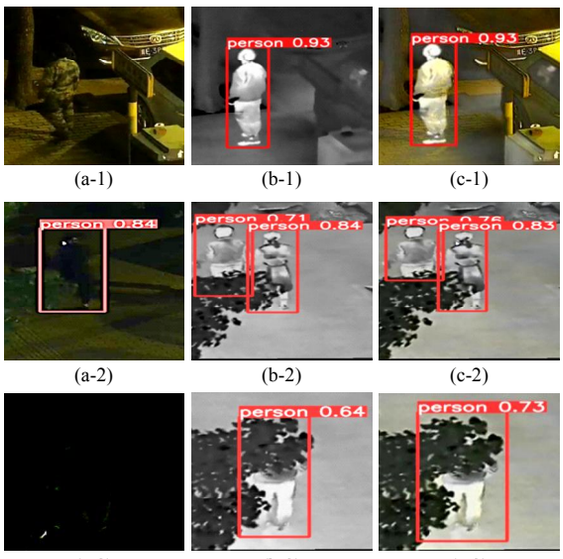


图7 不同模态图像的行人检测效果对比

Fig. 7 Comparison of pedestrian detection performance based on different-modal images

度和召回率分别达到94.7%和86.9%,表明本文通过优化模型结构,在减少漏检的同时抑制了误检;YOLOv5、YOLOv3-tiny、本文改进模型的模型权重均较小,但是本文改进模型在计算资源利用方面具有更高效率,其参数量在未改进YOLOv5模型的基础上减少32.1%。为同时保证行人目标检测的轻量化和精度要求,选取未改进YOLOv5模型和本文改进模型进行双模态图像检测效果对比分析,如图8所示,结果表明本文改进模型的置信度高于未改进YOLOv5模型的置信度,因此,本文改进模型能够有效地兼顾检测效率与精度要求。

表2 多种模型检测结果对比

Table 2 Comparison of detection results of different models

模型	精确率/%	召回率/%	mAP50/%	模型权重大小/MB
Faster R-CNN	91.3	81.3	90.3	108.2
Mask R-CNN	93.2	82.5	91.9	245.63
SSD	94.5	82.4	92.7	91.6
YOLOv3-tiny	90.4	81.0	89.1	17.4
YOLOv4	94.4	84.5	93.7	243.9
YOLOv5	95.2	86.2	94.2	14.0
本文模型	95.7	86.9	94.7	9.5

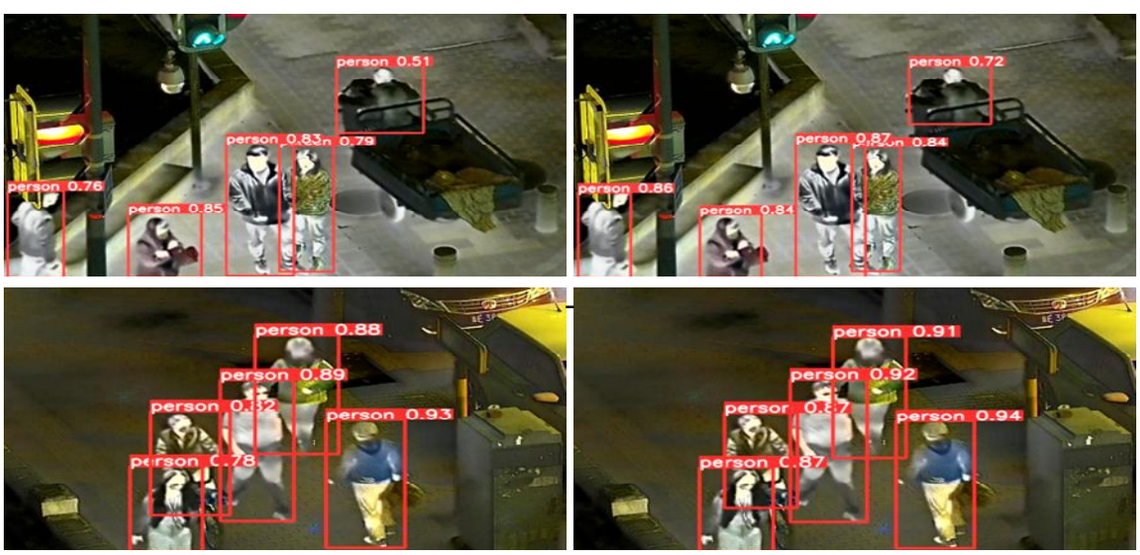


图8 未改进YOLOv5模型与本文改进模型的检测效果对比

Fig. 8 Comparison of detection effect between the unimproved YOLOv5 and the improved model

3.5 消融实验分析

通过消融实验验证本文改进模型中添加的不同模块对行人目标检测任务的性能提升是否确实有效。表3为GhostNet轻量化模块、GAM注意力机制模块、损失函数EIoU模块组合作用于本文改进模型的验证结果。

由表3可知:采用GhostNet模型将基线模型的权重大小从14.0 MB减少为7.4 MB,减少了47.1%,具有最小的模型权重和最高的帧率,实现了模型轻量化,但在检测精度上有所损失;加入GAM注意力机制使得模型的检测精度最高,超过95%,但是图像处理帧率最小;EIoU使得模型

表3　实验结果

Table 3　Experimental results

实验序号	Ghost Net	GAM	EIoU	mAP50/%	模型权重大小/MB	帧率/(帧·s^{-1})
1	—	—	—	94.2	14.0	115
2	√	—	—	92.5	7.4	120
3	—	√	—	95.3	15.9	97
4	—	—	√	94.5	14.4	116
5	√	√	√	94.7	9.5	119

的检测精度达到94.5%,且具有较高的图像处理帧率,但其模型权重较大;对于添加3个模块的本文模型,检测精度维持在94.7%,同时具备较小的模型权重和较高的帧率。实验结果表明,所采用的改进策略在保证检测精度的情况下,能够兼顾模型轻量化与检测精度。

4　结束语

(1)针对单模态可见光图像行人检测易出现误检、漏检的问题,本文采用SeAFusion算法对可见光与红外图像进行融合,提高了低照度环境下行人检测的鲁棒性。

(2)为适配行人实时检测场景,本文使用GhostNet替换YOLOv5原有主干网络,并在主干网络中引入GAM注意力机制,在降低模型复杂度的同时获得多维度特征,增强了混合模态图像行人识别能力,采用EIoU损失函数替换原有CIoU损失函数,能够实现更为准确的边界框回归和目标检测。

(3)改进后的模型权重大小降为9.5 MB,较未改进的YOLOv5模型权重降低32.1%,平均检测精度达到94.7%,且每秒检测的图像帧数相较改进前提高4帧。

(4)现有双模态数据集中的行人较为分散,未来将获取密集行人数据进行研究,进一步丰富行人感知应用场景,提升模型的工程应用价值。

参考文献:

[1] Liu J J, Zhang S T, Wang S, et al. Multispectral deep neural networks for pedestrian detection[C]// Proceedings of the British Machine Vision Conference, York, UK: BMVC, 2016.

[2] 郭晓晗, 彭理群, 马定辉. 基于车联网BSM数据与路侧视频融合的港口集装箱卡车碰撞危险辨识方法[J]. 交通信息与安全, 2023, 41(1): 1-12.
Guo Xiao-han, Peng Li-qun, Ma Ding-hui. A method of identifying collision risk of container trucks in port terminal areas under an integrated connected vehicle BSM and roadside video surveillance data[J]. Journal of Transport Information and Safety, 2023, 41(1): 1-12.

[3] 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): 47-58.
Zhao Bin, Wang Chun-ping, Fu Qiang, et al. Multiscale infraredpedestrian detection based on deep attention mechanism[J]. Acta Optica Sinica, 2020, 40(5): 47-58.

[4] Zhuang Y F, Pu Z Y, Hu J, et al. Illumination and temperature-aware multispectral networks for edge-computing enabled pedestrian detection[J]. IEEE Transactions on Network Science and Engineering, 2022, 9(3): 1282-1295.

[5] 游峰, 梁健中, 曹水金, 等. 面向多目标跟踪的密集行人群轨迹提取和运动语义感知[J]. 交通运输系统工程与信息, 2021, 21(6): 42-54, 95.
You Feng, Liang Jian-zhong, Cao Shui-jin, et al. Dense pedestrian crowd trajectory extraction and motion semantic perceptionbased on multi-object tracking [J]. Journal of Transportation Systems Engineering and Information Technology, 2021, 21 (6): 42-54, 95.

[6] Viola P, Jones M J. Robust real-time face detection [J]. International Journal of Computer Vision, 2001, 57(2): 137-154.

[7] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA: IEEE, 2005.

[8] 程旭, 宋晨, 史金钢, 等. 基于深度学习的通用目标检测研究综述[J]. 电子学报, 2021, 49(7): 1428-1438.
Cheng Xu, Song Chen, Shi Jin-gang, et al. A survey of generic object detection methods based on deep learning[J]. Acta Electronica Sinica, 2021, 49(7): 1428-1438.

[9] 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.

[10] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//IEEE International Conference on Computer Vision (ICCV), Venice, Italy: IEEE, 2017.

[11] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]//European Conference on Computer Vision. Amsterdam, The Netherlands: ECCV, 2016.

[12] Adarsh P, Rathi P, Kumar M. YOLO v3-Tiny: object detection and recognition using one stage improved model[C]//2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS). Coimbatore, India: IEEE, 2020.

[13] 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 Yan, Zhang Chong-yang, Tian Yong-hong, et al. An overview of deep learning based pedestrian detection algorithms[J]. Journal of Image and Graphics, 2022, 27(7): 2094-2111.

[14] 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.

[15] Guan D Y, Cao Y P, Liang J, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection[J]. Information Fusion, 2019, 50:148-157.

[16] Xue Y J, Ju Z Y, Li Y M, et al. MAF-YOLO: Multi-modal attention fusion based YOLO for pedestrian detection[J]. Infrared Physics & Technology, 2021, 118: 103906.

[17] Li C, Wang Y D, Liu X M. A multi-pedestrian tracking algorithm for dense scenes based on an attention mechanism and dual data association[J]. Applied Sciences, 2022, 12(19): 9597-9607.

[18] 孙文财, 胡旭歌, 杨志发, 等. 融合GPNet与图像多尺度特性的红外-可见光道路目标检测优化方法[J]. 吉林大学学报:工学版, 2024, 54(10): 2799-2806.
Sun Wen-cai, Hu Xu-ge, Yang Zhi-fa, et al. Optimization of infrared-visible road target detection by fusing GPNet and image multiscale features[J]. Journal of Jilin University(Engineering and Technology Edition), 2024, 54(10): 2799-2806.

[19] 刘子龙, 沈祥飞. 融合 Lite-HRNet 的 Yolov5 双模态自动驾驶小目标检测方法[J]. 汽车工程, 2022, 44(10): 1511-1520.
Liu Zi-long, Shen Xiang-fei. YOLOv5 dual-modeautomatic driving small target detection method combining Lite-HRNet [J]. Automotive Engineering, 2022, 44(10): 1511-1520.

[20] Falaschetti L, Manoni L, Palma L, et al. Embedded real-time vehicle and pedestrian detection using a compressed tiny YOLO v3 architecture[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(12): 19399-19414.

[21] 王宏志, 宋明轩, 程超, 等. 基于改进YOLOv5算法的道路目标检测方法[J]. 吉林大学学报:工学版, 2024, 54(9): 2658-2667.
Wang Hong-zhi, Song Ming-xuan, Cheng Chao, et al. Road object detection method based on improved YOLOv5 algorithm[J]. Journal of Jilin University (Engineering and Technology Edition), 2024, 54(9): 2658-2667.

[22] Tang L F, Yuan J T, Ma J Y. Image fusion in the loop of high-level vision tasks: a semantic-aware real-time infrared and visible image fusion network[J]. Information Fusion, 2022, 82: 28-42.

[23] Han K, Wang Y H, Tian Q, et al. Ghostnet: more features from cheap operations[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2020.

[24] Liu Y C, Shao Z R, Hoffmann N. Global attention mechanism: retain information to enhance channel-spatial interactions[EB/OL]. 2024-11-10.

[25] Jia X Y, Zhu C, Li M Z, et al. LLVIP: a visible-infrared paired dataset for low-light vision[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021.