基于重载外骨骼环境感知的红外与可见光图像融合

梁 超[1] 来跃深[1]
LIANG Chao LAI Yueshen

摘 要 重载外骨骼作为助力搬运器械，可有效增强人体搬运相关运动能力，降低人体损耗。目前，重载外骨骼多侧重于某些特定任务或环境，无法满足穿戴者“独立安全”的在未知环境下搬运行走的需求。因此，为了解决搬运环境下全天候的环境感知可能带来的光线不足、曝光以及障碍物重叠，而带来的纹理细节丢失和引入边缘伪影等常见问题，文章提出了一种基于双模态图像融合的工厂环境下目标检测算法。设计了基于 YOLOv8 算法架构的红外与可见光图像融合基础框架。构建了双通道图像融合模块，并在颈部网络层同时引入 ECA 注意力模块与 ADown 下采样机制，增强模型识别小目标的能力，同时减小模型的计算难度。实验结果表明，所提方法与单一基于 YOLOv8 的红外和可见光图像检测结果相比，在自制工厂数据集下，召回率提升 7.41 个百分点、mAP@0.5 提升 3.1 个百分点。相较于同类融合算法 MFEIF，mAP@0.5 提升了 4.95 个百分点。

关键词 YOLOv8；外骨骼环境感知；下采样机制；注意力机制；图像融合

doi：10.3969/j.issn.1672-9528.2025.01.034

0 引言

近年来，随着深度学习理论方法的迅速发展，可穿戴外骨骼系统利用神经网络来感知复杂的环境场景成为了可能。

环境感知模块是可穿戴外骨骼系统中的关键环节。实现对周围环境信息的全面感知是实现外骨骼设备安全性与智能性辅助前行的保障，也是完成准确决策与控制的前提条件。美国罗德岛大学的 Zhang 等人 [1-2] 通过布置于腰部的激光传感器获取环境数据，提取特征，通过决策树实现对平地、上升 / 下降楼梯的识别。Kleiner 等人 [3] 则通过毫米波实现障碍物的识别和测距。河北工业大学的张燕等人 [4-5] 使用安装在人体腰部的二维激光雷达，获取环境数据，利用凝聚分层聚类算法提取线性特征，然后使用有限状态自动机实现了平地、上 / 下斜坡和上 / 下楼梯 5 种地形 95.8% 的识别。上述研究中尽管在外骨骼的环境感知中取得一定的进展，但在恶劣条件下对于目标的检测依旧存在着进一步改善目标误检率、提高查全率的难点技术问题。

针对这一问题，本文基于 YOLOv8 算法网络，开展了双模态图像融合的外骨骼环境感知的目标检测算法研究。主要工作包括为：

（1）设计了多通道融合模块，为了提高小目标的检测精度，在主干网络融合了 ECA 注意力机制，同时为了解决多通道融合带来的参数量上升的问题，在主干网络受用 ADown 下采样机制替换卷积，在不影响精度与查全率的情况下，减少参数量，为其后续部署在最小系统上做准备。

（2）基于外骨骼行走区域需要，设计并收集全天候工厂环境下的障碍物的可见光以及对应的红外数据集；本文以可见光图像的标签作为标准，即红外和可见光共用一套标签系统，为算法的模型训练提供可靠的数据基础。

1 双通道图像融合检测基础框架

1.1 基于 YOLOv8 的双通道图像融合检测框架

目前，传统的目标检测算法存在对于光线不足、遮挡严重和抗干扰能力不足等问题，因此提出了基于 YOLOv8 的改进算法。首先，设计了多通道融合模块 Fusion modle，将原来的 3 通道输入转化为 6 通道输入，使得红外与可见光图像构成图像互补，深度挖掘红外与可见光的互补特性；其次主干部分的传统卷积使用下采样机制（anisotropic downsample，ADown）[6] 来替代，扩大模型的感受野，并使得模型更加的轻量化；最后在骨干网络的中间引用高效的通道注意力机制（efficient channel attention，ECA）[7]，确保模型更加关注有用的信息，降低干扰信息的权重，提高融合质量，有效提高模型的检测能力，经过上述改进，算法的各个指标均有明显提升，改进后的模型如图 1 所示。

1. 西安工业大学机电工程学院 陕西西安 710021
[基金项目] 陕西省自然科学基础研究计划项目（2021JM-020）

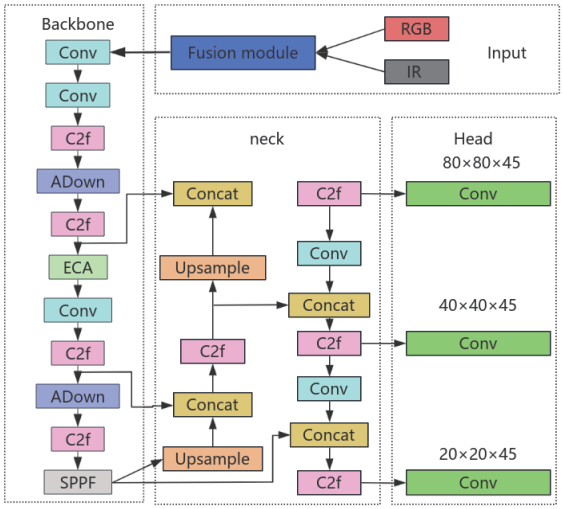

图 1 改进后的结构图

1.2 Fusion module

基于红外与可见光图像形成互补图像对，特征图中各位置的特征权重随着互补特性而变。为了深入利用红外与可见光图像的互补性质，本文构建了输入特征融合模块 Fusion module，用以实现红外与可见光的自适应融合，其模型结构如图 2 所示。

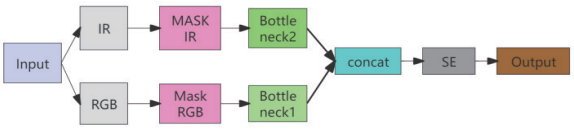

图 2 Fusion module 融合模块

首先，输入进来的红外与可见光图像在通道维度上使用 Sigmoid 函数生成红外与可见光图像的权重，因为红外和可见光具有互补的特性，因此设红外权重为 q，则可见光的权重为 $1-q$，计算公式为：

$$q=\sigma(B_i(G_r)) \tag{1}$$

式中：q 代表红外权重；$\sigma(\cdot)$ 代表 Sigmoid 函数；$B_i(\cdot)$ 代表批处理的结果；G_r 代表全局特征。

获得红外与可见光的权重之后，使得融合特征同时具备全局特征与局部特征，计算公式为：

$$G_{ir}=I_{ir}\oplus(I_{rgb}\otimes(1-q)) \tag{2}$$

$$G_{rgb}=I_{rgb}\oplus(I_{ir}\otimes q) \tag{3}$$

式中：$\otimes$代表乘积；$\oplus$代表求和；G_{ir}、G_{rgb} 处理后的红外与可见光特征图；I_{ir}、I_{rgb} 代表输入的红外与可见光特征图；i 代表使用频次；$C(\cdot)$ 代表 Concat 函数。

其次，对处理后的红外与可将光图像的特征图进行拼接用于捕捉全局特征，之后利用卷积平滑处理产生更加精准的信息，计算公式为：

$$G_{r1}=\mathrm{Conv}^2(\mathrm{AAP}(C(F_{ir}+F_{rgb}))) \tag{4}$$

式中：G_{r1} 代表处理后的全局特征；$\mathrm{Conv}^i(\cdot)$ 代表点卷积层；$\mathrm{AAP}(\cdot)$ 代表自适应平均池化函数。

1.3 ECA 注意力机制

通道注意力机制的目标是自适应地调整通道特征的权重，使得网络可以更好地关注重要特征，抑制不重要特征。ECA 在卷积操作中引入通道注意力机制，以捕捉不同通道之间的关系，从而提升特征表示能力，其结构如图 3 所示。

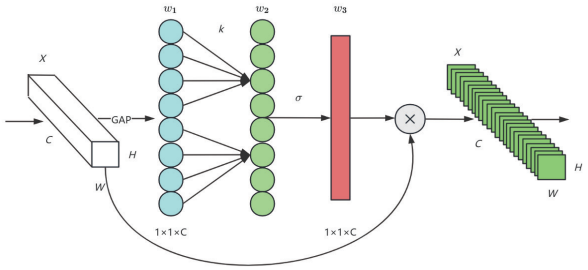

图 3 ECA 注意力机制机构图

首先，将输入 X 通过全局平均池化（globel average pooling，GAP），得到 $1\times1\times C$ 的特征向量 $\boldsymbol{\omega}_1$，实现全局上下文信息融合；然后计算自适应卷积核 k 的大小，经过卷积核大小为 k 的一维卷积得到一个通道权重向量 $\boldsymbol{\omega}_2$；再经过 Sigmoid 激活函数将权重向量 $\boldsymbol{\omega}_2$ 映射在 0~1 之间得到 $\boldsymbol{\omega}_3$；最后，将 $\boldsymbol{\omega}_3$ 与输入特征图 X 相乘，获得加权后的特征图 X。

ECA 能够避免通道降维造成的损失，在不增加过多参数和计算成本的情况下，有效地增强网络的表征能力。

1.4 ADown 下采样

为了改善多通道会带来的数据量增加的问题，采用 ADown 的下采样方式通过最大池化与平均池化的结合，平均池化保留全局信息，最大池化有助于捕捉局部的特征信息，两种池化相结合，实现不同尺度下的特征提取。ADown 下采样机制的采样过程如图 4 所示。

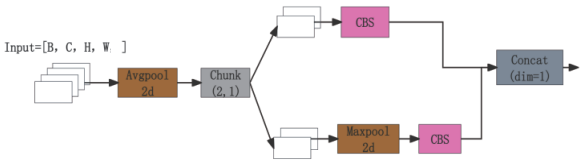

图 4 ADown 下采样模型结构图

首先在输入特征图上滑动一个固定大小的窗口，计算窗口中所有像素的平均值，并将该平均值作为输出的相应位置的像素值。这个过程可以通过下采样来实现，从而减少输出特征图的尺寸，同时保留主要的特征。有助于减少模型的参数数量和计算量，并且可以增强模型的平移不变性，使其对输入数据的微小变化具有更好的鲁棒性。然后沿着通道维度 C 将张量分割成多个块，每个块包含两个通道。从而将模型的通道分组，以便于处理不同通道之间的关系或增强特征提

取能力。上一步操作所分的两个通道为 CBS（convolutional block softmax）和 Maxpool2d。再接一个 CBS 模块。最后将两个分支的结果进行拼接，不仅降低了参数量和计算量并且最小化分割图与真实标签之间的差异。通过引入多分支结构提供了更多的特征组合和信息交互，保留更多的上下文信息，从而防止一些重要特征丢失过多。

1.5 评估指标

本文采用精确率（Precision）、召回率（Recall）以及平均精度均值（mAP）作为评价标准，采用的计算公式为：

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

$$\text{mAP} = \frac{1}{n}\sum_{i=1}^{n} \text{AP}_i \tag{6}$$

式中：TP 为正确检测到的目标数量；FN 为未检测到的目标数量。

2 实验与结果分析

2.1 实验环境及参数设置

本实验的硬件配置如下：CPU 为 22 vCPU AMD EPYC 7T83 64-Core Processor、内存 90 GB、显卡型号为 NVIDIA RTX 4090(24 GB)。模型框架采用 PyTorch1.11，Python3.8，CUDA11.3，操作系统为 Ubuntu 20.04。

2.2 数据集

本文采用自制的工厂数据集对算法进行评估。该数据集包含了 1920 对红外和可见光图像，其中包含训练集 1520 对图像，验证集包含 810 对图像。数据集包含 6 个类别，为工厂中可能出现在外骨骼穿戴者行进路线上最可能出现的障碍物，分别为：“box”“trolleys”“chair”“line1”“line”“people”。“box”为公司内部大量的可移动机箱；“trolleys”为装载货物的小推车，会经常移动，出现在各个位置；“chair”为椅子，工人需要在各个车间地方工作，椅子会经常移动；“line1”为直的线缆，因为公司存在大量的接电设备，以及调试设备等工作，存在大量线缆，为了方便识别；“line”划分为曲线，盘起来的线缆，最后一个类别为人，会出现在各个地方。为了训练中仅存在一种变量，本文以可见光图像标签作为标准，即红外与可见光共用一套标签系统，然后将红外图配准至可见光图像上，方便提取同名点，提升数据集的精度与合理性。具体的标签数量为“box”有 1740 个，“trolleys”有 547 个，“chair”有 361 个，“line1”有 423 个，“line”有 329 个，“people”有 535 个。

2.3 消融实验

为了应对工厂环境下复杂恶劣的光线环境，本文在自制的工厂数据集进行了消融实验，为验证各个改进的重要性。将每个改进依次嵌入到 YOLOv8 的模型之中，并且使用相同的训练参数和环境条件，实验结果如表 1 所示。

表 1 消融实验

实验	FM	ECA	ADown	*R*/%	*P*/%	mAP@0.5/%
1				79.2	83.22	86.3
2	√			86.5	78.69	88.5
3	√	√		86.2	77.65	87.8
4	√		√	85.2	80.2	88.8
5	√	√	√	86.6	79.29	89.4

表 1 中的实验 2~5 为了评估不同的改进策略对于 YOLOv8 的影响。实验 1 为原始的 YOLOv8 在自建数据集上的训练结果；实验 2 为在输入端加入 FM 融合模块，结果可以看到，召回率、mAP@0.5 有明显提升；实验 3 为在输入端加入融合模块以及主干网络加入注意力机制，可以发现的是因高效通道注意力的增加，推理速度加快；实验 4 为添加的 FM 融合模块和 ADown 下采样模块，可以看到，参数量明显下降，以及精度和召回率有不同程度的提升，以方便后续模型的部署；实验 5 是将 3 个改进同时融合，可以发现指标均有明显提升，召回率提升 7.41 个百分点，mAP@0.5 提升 3.1 个百分点，各方面性能均有提升。

综上所述，实验数据清晰地验证了这些改进点共同提高了 YOLOv8 在工厂搬运环境下目标检测方面的性能，降低了漏检的情况，达到了实验目的。

2.4 不同算法对比试验

为了验证本文方法在双模态图像检测领域的性能以及算法的泛化性，本文引入了 3 种当前融合检测网络较好的方法进行对比验证，3 种算法分别为 SwinFusion [8]、MFEIF [9]、NestFuse [10]。表 2 为在自建数据集上不同双模态图像融合检测方法对比实验的结果。

表 2 对比实验

算法模型	*R*/%	*P*/%	mAP@0.5/%
YOLOv8	79.20	83.22	79.24
SwinFusion	80.31	71.26	80.58
NestFuse	79.60	70.28	82.34
MFEIF	83.27	76.25	81.66
改进算法	86.61	79.29	86.61

从表 2 中可得，SwinFusion、MFEIF、NestFuse 算法的检测结果均优于单模态图像检测结果，验证了双模态的融合方式能够有效提升目标的检测性能。同时，相比于对比方法中的检测效果最好的 MFEIF 算法，本文算法在召回率、准确率、mAP@0.5 均有所提升，分别提升了 3.34、3.04 和 4.95 个百分点，说明本文方法的性能由于对比方法，能够适应于双模态的融合检测场景。

2.5 模型泛化性实验

本文引入 LLVIP 行人目标检测数据集，进行通用性实验，以证明本文方法的泛化能力。

LLVIP 行人目标检测数据集为处理过的数据集，原数据集包含 30 976 张红外与可见光图像，共计 15 488 对。为了更快地、进行测试验证，从中筛选了 1000 张红外与可见光图像，按照 8:2 划分为训练集和验证集。将筛选出来的图像进行手工标注，仅设“person”类别。共标注有 2016 个标签，其中训练集有 1621 个标签，验证集有标签 432 个标签。

在 LLVIP 行人目标检测数据集中，本文做了融合检测框架对比实验和双模态图像融合 检测方法对比试验。实验结果如表 3 所示。

表 3 对比实验结果

算法模型	*R*/%	*P*/%	mAP@0.5/%
YOLOv8	66.21	77.54	66.21
改进算法	78.71	87.93	78.71

由表 3 可知，改进的双模态融合算法的性能明显优于单模态图像的检测方式，同样证明了图像融合方式能够有效提高目标检测精度，并且能够适应更加恶劣的环境。在 LLVIP 数据集上召回率 *R*、准确率 *P*、mAP@0.5 分别提升了 12.5、10.39 和 12.5 个百分点，证明了在 LLVIP 数据集中，改进的多模态融合算法具备更高的评估结果，显著提高了目标检测的精度，证明其具有良好的泛化能力。

3 结论

针对目前外骨骼环境感知部分中的障碍物检测方面，存在着环境恶劣的光线条件，使得视觉检测出现错检以及漏检的问题，开展了红外与可见光图像融合检测算法研究，并对 YOLOv8 目标检测网络进行改进，为了使其更好的捕捉小目标的信息，在颈部网络引入高效通道注意力机制 ECA；同时为了降低多模态检测带来的参数量提升，使用 ADown 下采样机制，降低模型的参数量，使得模型更加的轻量化，更便于后续的算法部署。最后构建了多模态融合模块 Fusion Medole，利用互补图像的权重图来融合双模态图像的特征信息，进一步提升模型的检测精测精度。实验结果表明，无论是在曝光、暗光环境下，或者是模糊和目标部分被遮挡等复杂环境下，本文方法均表现出出色的检测性能，在工厂环境下的目标检测具有更高的适应性，为复杂的工厂环境下的外骨骼环境感知目标检测提供了有力支持。

参考文献：

[1] ZHANG F, FANG Z, LIU M, et al. Preliminary design of a terrain recognition system[C/OL]//2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. Piscataway: IEEE, 2011[2024-03-16].https://ieeexplore.ieee.org/document/6091391.

[2] 刘俊明 , 孟卫华 . 基于深度学习的单阶段目标检测算法研究综述 [J]. 航空兵器 , 2020, 27(3): 44-53.

[3] KLEINER B, ZIEGENSPECK N, STOLYAROV R, et al. A radar-based terrain mapping approach for stair detection towards enhanced prosthetic foot control[C/OL]// 2018 7th IEEE International Conference on Biomedical Robotics and Biomechatronics (Biorob). Piscataway: IEEE, 2018[2024-05-10].https://ieeexplore.ieee.org/document/8487722.

[4] 张燕 , 许京 , 陈玲玲 , 等 . 基于激光距离传感器的路况识别系统的设计 [J]. 激光与红外 , 2016, 46(3): 265-270.

[5] 任钰 . 基于 faster R-CNN 的小目标检测研究与应用 [D]. 安庆 : 安庆师范大学 , 2022.

[6] WANG C Y, YEH I H, LIAO H Y M. YOLOv9: learning what you want to learn using programmable gradient information[DB/OL]. (2024-02-21)[2024-03-14].https://doi.org/10.48550/arXiv.2402.13616.

[7] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C/OL]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020[2024-04-05].https://ieeexplore.ieee.org/document/9156697.

[8] MA J Y, TANG L F, FAN F, et al. SwinFusion: cross-domain long-range learning for general image fusion via swin transformer[J]. IEEE/CAA journal of automatica sinica, 2022, 9(7): 1200-1217.

[9] LIU J Y, FAN X, JIANG J, et al. Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion[J]. IEEE transactions on circuits and systems for video technology, 2021, 32(1): 105-119.

[10] 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.

【作者简介】

梁超（2000—），男，陕西商洛人，硕士，研究方向：计算机视觉。

来跃深（1965—），男，陕西西安人，博士，教授，研究方向：计算机控制与测量、机器智能控制等。

（收稿日期：2024-10-11）