分类号：TP391　　学校代码：10363
密　级：公开　　学　　号：2220320186

安徽工程大学

Anhui Polytechnic University

硕士学位论文

题目　面向夜间道路场景的行人检测研究

论 文 作 者　凌成
校 内 导 师　许钢（教授）
校 外 导 师　朱标（高级工程师）
专业学位类别　电子信息
研究方向（领域）　控制工程

论文提交日期：2025 年 5 月 19 日

分类号：TP391　　单位代码：10363

密　级：公开　　学　　号：2220320186

题　　目　面向夜间道路场景的行人检测研究

题　　目　**Research on Pedestrian Detection for Nighttime Road Scenes**

学 生 姓 名：凌 成

校 内 导 师：许 钢（教 授）

校 外 导 师：朱标（高级工程师）

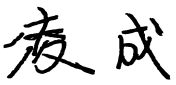
专　　业：电子信息

研 究 方 向：控制工程

论文答辩日期：2025 年 05 月 12 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名： 

日期：2025 年 05 月 12 日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 □，在 ____ 年解密后适用本版权书。

本学位论文属于

不保密 □。

学位论文作者签名  指导教师签名： 

日期：2025 年 05 月 12 日　　日期：2025 年 05 月 12 日

面向夜间道路场景的行人检测研究

摘 要

行人检测是目标检测领域的重要研究方向，在智能驾驶技术快速发展的背景下具有关键作用。现有研究多聚焦于光照条件良好的场景，而对低可见度场景下的行人检测问题关注不足。由于在夜间道路场景中可见光相机易受光照限制导致成像质量下降，智能驾驶系统对周围环境的感知能力显著降低，存在较大的交通隐患。在此背景下，红外热成像技术因其不受环境光照影响特性成为有效的解决方法。因此，本文基于夜间道路场景对行人检测展开相关研究，主要工作如下：

（1）构建符合本文研究场景的夜间红外图像数据集。由于开源数据集包含场景众多，因此首先基于 FLIR 数据集的夜间道路场景子集分别构建夜间红外图像数据集、夜间跨模态图像数据集作为本文算法验证的数据集。其次针对红外图像低对比度的特点，采用数字细节增强(Digital Detail Enhancement, DDE)算法提升图像质量，最后验证图像各个指标的有效改善，解决现有数据集场景混杂、质量不均的问题。

（2）针对夜间可见光成像效果不佳的道路场景，提出轻量化红外行人检测模型 RT_YOLOv8。针对现有模型在夜间道路场景存在的计算冗余、特征适应性弱问题，设计四阶段改进方案：首先采用 RepGhostNet 作为主干网络，提升整体行人特征提取能力；其次引入 Triplet 注意力机制，通过宽度、高度、通道三维度的交互增强目标特

征表达能力；第三，利用可变形卷积(Deformable Conv v2, DCNv2)重构 C2f 模块以增强空间适应性；最后在 Detect 模块中采用 PConv 替换部分常规卷积，减少冗余计算并提升空间特征提取效率。实验表明，RT_YOLOv8 相较于 YOLOv8 在实验建立的红外数据集上平均精度提升 0.5%，同时计算量(GFLOPs)降低 50.62%。

（3）针对夜间低照度复合的道路场景，单一采用红外图像具有一定局限性，提出基于双分支中期融合的网络架构，设计五阶段改进方案：首先构建基于 YOLOv8 的双分支网络进行特征提取；其次在进入全卷积层前采用中期融合策略实现双模态特征融合；第三，在 SPPF 模块中融入大核可分离注意力机制(Large Separable Kernel Attention, LSKA)，在不增加计算量的前提下扩展感受野；第四，集成 RT_YOLOv8 核心模块(C2f_DCNv2、Triplet attention、Detect_PC)以维持轻量化特性；最后通过引入 MPDIoU 损失函数优化基准网络来提升检测框定位精度和模型收敛速度。实验表明，该跨模态融合算法相较单红外和可见光模态平均精度分别提升 1.5%、2.2%。

关键词：夜间行人检测，红外图像，轻量化，跨模态，YOLOv8

RESEARCH ON PEDESTRIAN DETECTION FOR NIGHTTIME ROAD SCENES

ABSTRACT

Pedestrian detection is an important research direction in the field of object detection, which plays a key role in the context of the rapid development of intelligent driving technology. Most of the existing studies focus on the scene with good lighting conditions, but not enough attention is paid to the problem of pedestrian detection in the low visibility scene. Due to the deterioration of imaging quality caused by the limitation of light in the night road scene, the perception ability of the intelligent driving system to the surrounding environment is significantly reduced, and there are great traffic hazards. In this context, infrared thermography has become an effective solution because it is not affected by ambient light. Therefore, this paper conducts relevant research on pedestrian detection based on night road scenes, and the main work is as follows:

(1) A nighttime infrared image dataset that conforms to the research scenario in this paper is constructed. Due to the large number of scenarios in the open-source dataset, the nighttime infrared image dataset and the nighttime cross-modal image dataset were constructed based on the subset of the night road scene of the FLIR dataset as the datasets verified by the algorithm in this paper. Secondly, aiming at the characteristics of low contrast of infrared images, the Digital Detail Enhancement (DDE) algorithm is used to improve the image quality, and finally the effective improvement of each index of the image is verified to solve the problems of mixed scenes and uneven quality of the existing datasets.

(2) Aiming at the road scene with poor visible light imaging effect at

night, a lightweight infrared pedestrian detection model RT_YOLOv8 is proposed. In order to solve the problems of computational redundancy and weak feature adaptability of the existing models in night road scenes, a four-stage improvement scheme was designed: firstly, RepGhostNet was used as the backbone network to improve the overall pedestrian feature extraction ability; Secondly, the Triplet attention mechanism was introduced to enhance the expression ability of target features through the interaction of width, height and channel. Thirdly, the deformable convolution (DCNv2) was used to reconstruct the C2f module to enhance the spatial adaptability. Finally, PConv is used to replace part of the conventional convolution in the Detect module to reduce redundant calculations and improve the efficiency of spatial feature extraction. Experiments show that compared with YOLOv8, the average accuracy of RT_YOLOv8 is increased by 0.5% on the infrared dataset established by the experiment, and the computational cost (GFLOPs) is reduced by 50.62%.

(3) In view of the road scene of low illumination at night, the single use of infrared images has certain limitations, and a network architecture based on dual-branch mid-term fusion is proposed, and a five-stage improvement scheme is designed: firstly, a dual-branch network based on YOLOv8 is constructed for feature extraction; Secondly, the medium-term fusion strategy is used to achieve dual-modal feature fusion before entering the fully convolutional layer. Thirdly, the Large Kernel Separable Attention Mechanism (LSKA) was integrated into the SPPF module to expand the receptive field without increasing the amount of computation. Fourth, integrate RT_YOLOv8 core modules (C2f_DCNv2, Triplet attention, Detect_PC) to maintain lightweight characteristics; Finally, the MPDIoU loss function is introduced to optimize the benchmark network to improve

the positioning accuracy of the detection frame and the convergence speed of the model. Experiments show that the average accuracy of the cross-modal fusion algorithm is 1.5% and 2.2% higher than that of single infrared and visible light modes, respectively.

Ling Cheng(Electronic Information)

Supervised by Professor Xu Gang

KEY WORDS: Night pedestrian detection, Infrared image, Lightweight Cross-modality, YOLOv8

目录

第1章 绪论

目标检测作为计算机视觉领域的核心课题，通过深度学习技术来实现图像高层次语义信息理解的重要基础。与人类视觉系统具备的先天感知能力不同，计算机对数字图像无非是红、绿、蓝三原色的集合。目标检测包含两个关键子问题：其一为定位(Localization)，即通过边界框精确标定目标在环境中的具体位置；其二为分类(Classification)，即准确识别目标所属的语义类别。

近些年来，随着深度学习理论的发展，特别是卷积神经网络(Convolutional Neural Networks, CNN)与 Transformer 架构的突破性进展，将图像处理运用到生活中已成为时代的趋势。深度学习的快速发展推动了图像目标检测急速的提升，从检测行人拓展到形形色色的目标。值得关注的是，目标检测技术的实用化进程已突破传统算力限制，基于神经网络剪枝、知识蒸馏等模型压缩方法，部分算法在边缘计算设备(如 Jetson Nano)上可实现实时检测。为构建通用视觉感知系统提供了新的可能性。

1.1 研究背景与意义

计算机视觉是人工智能领域的核心领域之一，通过图像传感器与深度学习的协同工作，模拟人类视觉的运作机制。随着深度学习算法的突飞猛进，该技术已经实现在智能驾驶、智能安防大规模部署。以智能驾驶[1]为例，目前智能驾驶系统按照自动化程度可分为两类，一类为 L4 级自动驾驶，通过降低人力成本来为用户带来更好的乘车体验，如 2024 年 Tesla 发布的无人驾驶出租车 Cybercab，如图 1-1 所示，其与传统车辆相比，亮点在于减少了方向盘和踏板。另一类为 L2-L3 级辅助驾驶系统，简单来说，就是汽车高级辅助驾驶系统(Advanced Driver Assistance System, ADAS)，ADAS 通过定位、感知、预测、规划和控制来减少人的参与，不仅提高了汽车驾驶安全性还提高了道路利用率来减少道路拥堵。在现代化城市交通运行中，在人为驾驶汽车时，通常通过眼睛来判断行驶速度、车辆位置和转弯角度。在智能驾驶中，主要通过车载静态摄像头结合计算机视觉算法实现环境感知，同时配合毫米波雷达、激光雷达等多种传感器融合技术来代替人

眼去感知环境，从而进一步对汽车发出行驶指令。通过计算机视觉获取道路信息后，结合 V2X 车联网技术将实时交通数据上传至云端。联合高精地图、导航算法等其他技术协同，最终由路径规划算法生成优化路线，在此期间使用大量计算机视觉技术来保证智能驾驶的安全性。

图 1-1 Tesla 无人驾驶出租车 Cybercab

2020 年，国家发改委联合工信部等部委共同发布《智能汽车创新发展战略》，确立多维度发展：

智能化路线：2025 年完成 L3 级自动驾驶产业化应用，推动 L4 级自动驾驶在限定场景下的商业化应用。

网联化布局：2025 年规划建成 LTE-V2X(Vehicle-to-Everything)区域覆盖，实现 5G-V2X 在典型城市与高速公路的示范运行，达到高精度时空基准服务网络全覆盖的目标。

标准体系建设：2025 年前构建完成符合中国技术路线的智能汽车标准框架。

参照美国国家公路交通安全管理局(National Highway Traffic Safety Administration, NHTSA)技术路线，汽车智能驾驶可以分为四个阶段。如图 1-2 所示。

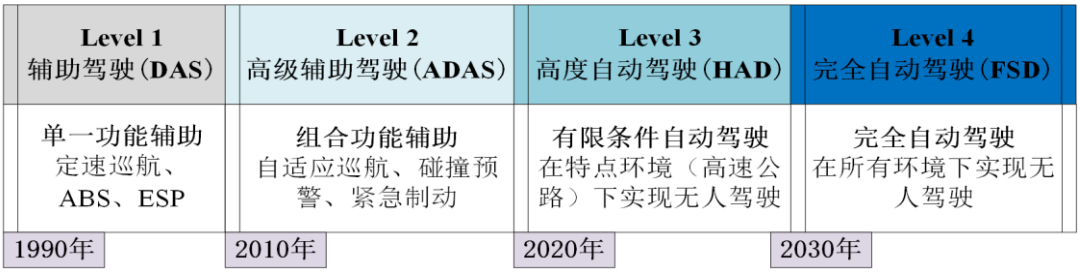


图 1-2 智能驾驶阶段图

行人保护系统(Pedestrian Protection Systems, PPS)是 ADAS 的重要的组成部分，行人保护系统负责对车辆周围区域进行行人检测，保证在智能驾驶过程中及时地避开行人，其核心为行人检测系统(Pedestrian Detection Systems, PDS)，在城市主干路对于其技术要求更加高，要求其在保证系统实时性的同时提高准确性。行人检测作为计算机视觉的重要研究方向，在其初期就备受关注，并且在近些年得到了大量的技术支持及应用，行人检测领域大量工程师提出大量目标检测算法。

然而当前普遍行人检测算法所针对的场景是基于可见光条件良好的情况下的，实际应用场景复杂多变，尤其在夜间可见光严重不足甚至无可见光的场景比比皆是。根据世界卫生组织统计，全球夜间交通事故率较白天高出 3 倍，夜间场景不仅是不可忽视的，反而是行人检测领域所需研究内容的重中之重。在可见光较弱甚至无可见光的情况下，普通摄像头的成像效果则会出现较差甚至无法成像的可能性，进一步会导致在视觉传感器识别目标难以提取到有效的目标信息，此时普通摄像头检测功能就相形见绌了，需要一种可以在恶劣环境中能更好地对目标进行成像的设备。红外热成像[2-5]的出现在一定程度上可以很好的缓解上述问题，由于人体具有比周围环境更高的温度，其热辐射可以更好被红外相机捕捉到。红外热成像具有不依赖光照条件等优点。由于红外图像为单通道灰度图像，缺乏颜色和纹理信息，导致基于深度学习的检测模型难以提取其中高区分度的特征，所以在图像中目标的细节轮廓的描述上也具有一定的缺陷。

在本文的研究内容中，主要基于深度学习框架，针对单模态（红外图像）和跨模态（红外图像+可见光图像）融合的行人检测算法进行改进，进一步提高行人检测的精度并且降低算法的模型大小。主要针对基于夜间红外图像场景下的行人检测和基于夜间场景下红外图像和可见光图像跨模态的行人检测展开相关研究。

1.2 国内外研究现状

近 20 年来，行人检测技术的发展日新月异，从早期的技术萌芽到如今的体系化成熟，已完成了从实验室原型到产业化应用的完整蜕变。国内外众多学者深耕于此领域，主要包括针对可见光图像和红外图像开展相关的课题研究与项目。从研究方向上看，主要包括基于特征提取、基于深度学习等方法。目标检测是计

算机视觉的一个核心方向，主要任务是对于所捕获的物体进行识别和分类。近些年来由于深度学习的崛起，神经网络可以在大量的数据中筛选出有用的信息并且加以学习使得具有较强的特征提取和拟合能力，继而出现大量优秀的目标检测算法。当前基于深度学习的红外检测算法有两类，一类是双阶段目标检测算法[6-8]，在早期研究人员主要围绕分类问题进行研究实验，首先根据图像识别目标提取目标框，然后对目标框分别进行多次修正，然而这样带来的效果则是检测速度较慢。出现 R-CNN(Region-CNN)[9-11]，随之而来 Fast R-CNN[12-14]、Faster R-CNN[15]等算法表现出优良的性能，所以直至当前 Faster R-CNN 仍然是目标检测的常青树。另一类是单阶段目标检测算法，单阶段目标检测算法是对图像进行直接的检测再生成检测结果，所以这种方式带来的弊端显而易见，导致目标检测精度较低。目前具有此类典型算法有 YOLO[16-18](You Only Look Once)列算法和 SSD[19-20](Single Shot MultiBox Detector)系列算法。

1.2.1 红外图像行人检测研究现状

随着非制冷型红外焦平面阵列技术的突破，红外图像行人检测研究在近五年取得显著进展。根据国际光学工程学会(Society of Photo-Optical Instrumentation Engineers, SPIE)2023 年发布的技术研究报告显示，当前主流基于红外图像的目标检测算法可以划分为以下三大技术路线：基于机器学习的决策方法、基于阈值分割的方法、基于兴趣区域搜索的方法。红外图像的行人检测算法发展历程与目标检测发展历程基本一致，都经历了传统手动提取特征到深度学习方法提取特征的过程。

2001 年，Viola P, Jones M[21]提出一种基于自适应增强算法(Adaptive boost, AdaBoost)和一种复杂的分类器组合成级联的方法，特定于待测对象的注意力机制算法。还引入一种名为“积分图像”的新图像表示，可以快速的实现特征计算，从而保证了检测的实时性与准确性的兼顾。

2007 年，Zhang L[22]等人将描述人体轮廓信息的“小边”(Edgelet)和方向梯度直方图特征应用于红外图像行人轮廓的描述。还利用自适应增强算法(Adaptive Boost, AdaBoost)从数据中筛选出一组效果最好的“小边”，用于构造头部、四肢的分类器，再利用联合似然模型构建可重构完整人体表征的分类模型，通过协同

优化红外热成像数据特征与可见光光谱，成功实现了红外图像行人检测的实时性技术的提高。

2010 年，Olmeda[23]等人在红外特征提取领域取得突破性进展，提出一种基于相位一致性原理的局部能量方向直方图(Histograms of Oriented Phase Energy, HOPE)来描述红外行人特性。该特征通过量化局部能量场的相位分布特性，有效增强了红外图像中行人目标的边缘响应，其背景鲁棒性显著优于传统检测方法，在复杂场景中展现出更优的检测精度。

2009 年，Wang X[24]等人为进一步提升局部二值模式(Local Binary Patterns, LBP)在人体检索中的表征局限性，创新性地构建了层次化特征编码框架。该研究引入空间分块策略，将检测窗口离散化为等尺寸单元网格，通过单元级 LBP 直方图计算与特征级联融合，构建复合型纹理描述算子。这种类 HOG(Histogram of Oriented Gradients)的层级特征编码机制有效增强了人体局部纹理特征的表征粒度，相较于传统全局 LBP 方法展现出更优的检索鲁棒性。2014 年 Mi[25]等人在局部二值模式(LBP)理论框架基础上提出一种融合能量对称与中心对称双约束的定向中心对称的局部二进制模式(Oriented Center-Symmetric Local Binary Patterns, OCS-LBP)的行人检测算法，该算法通过双重约束机制重构局部特征编码空间，利用中心对称核函数增加梯度方向敏感性，显著优化了行人目标微结构的特征表征能力，在复杂场景下表现出优于传统 LBP 的检测鲁棒性。

2018 年，刘峰[26]等人构建了层次化特征融合的级联检测框架，创新性地将形态学约束与特征工程相结合。该模型首先建立基于尺寸比例先验的 Haar 特征筛选器，实现红外图像中行人待检测区域快速提取；继而通过设计动态尺度空间压缩机制，耦合改进型梯度直方图，通过子空间投影优化显著降低特征维度；最后构建支持向量机[27](Support Vector Machine, SVM)驱动的多核学习分类器，在特征融合层实现非线性决策边界建模。

王晓红[28]等人提出一种针对较远距离小尺度行人的 YOLOv5-p4 夜间行人检测模型，通过增加更小目标的检测层，引入 BIFPN 特征融合机制，使检测网络更好聚焦于物体细节特征。代少升[29]等人提出一种基于视觉 Transformer 和双解码器的红外目标检测方法，使用视觉 Transformer 作为编码器，从而提取目标图像的多尺度特征。设计一个由交互式解码器和辅助解码器组成的双解码器模块，

通过双解码器模块能够充分利用不同特征之间的互补信息，促进深层特征与浅层特征之间的交互，并通过双解码器的结果进行叠加，提高解码器对红外目标的重构能力。刘学[30]等人提出一种改进 SSD 红外行人检测模型，通过利用深度可分卷积方法降低特征提取网络模型大小，在网络中加入 SENet 注意力机制模块，重新分配各个特征提取通道权重，提升算法对行人目标检测的针对性。

1.2.2 跨模态图像融合行人检测研究现状

图像融合的目的是将同一位置不同场景下的图像通过常见的融合方法融合成一张图片，融合后的图像具有原始图像的固有特征，通过互补来弥补原始图片的缺陷。当前可融合的图像类型众多，如合成孔径雷达(Synthetic Aperture Radar, SAR)图像与可见光图像、电子计算机断层扫描(Computed Tomography, CT)图像与核磁共振成像(Nuclear Magnetic Resonance Imaging, NMRI)图像、红外图像与可见光图像等。其中红外图像可以很好避免外界环境因素对其带来的影响，具有很好的夜间分辨能力，可见光图像具有充分的场景特征与高分辨。然而其都具有各自的劣势，红外图像为灰度图像，其分辨率较可见光图像则不值一提，且其背景特征较为模糊，可见光图像由于侧重度不同，无法很好将图像中行人剥离出来，并且在可见光不足的场景下，其采集能力就大大降低。当然如果可以很好将二者结合起来，取其之长补彼之短则显得更为重要。为了解决上述问题，Hwang[31]等人结合红外图像与可见光图像进行目标检测，实验表明基于红外-可见光图像多模态相较于单模态其检测精度具有较大提升，并构建一个多模态图像数据集，促进了相应的科学研究[32-40]。

图像融合技术依据处理层级差异主要可以划分为三个类别：像素级图像融合、决策级图像融合、特征级图像融合。

（1）像素级图像融合：该融合方法是在像素层面对图像进行融合，可以保留较多的图像细节，属于低层次的融合。常见的像素级融合有加权平均法、主成分分析法(Principal Components Analysis, PCA)、多分辨率融合。H.Bhujle[41]等人提出一种多波段图像融合策略，由于图像融合时是使用相应位置像素的标准化加权平均值计算的，为了获取权重，在图像中寻找到富含重要局部细节的差异。为

了增加具有高权重值的有效像素的数量，使用双边核进行非局部均值滤波器，进一步实现可分离的非局部均值滤波器可大大降低计算成本。

（2）决策级图像融合：该融合方法通过对源图像进行目标识别后划分类别，再对划分类别后的结果进行决策级的融合，得到最后结果，属于高层次的融合。常见的决策级融合有：基于神经网络的融合、基于支持向量机的融合。Zhao S[42]等人提出一种基于 SVM 的图像决策级融合新方法和相应的基于共识理论的融合规则，有效提高了融合图像的整体分类准确率。

（3）特征级图像融合：该融合方法作为中间层次的处理方式，采用特征提取技术对原始图像进行信息解构，重点针对目标物体的几何特征（如边界形态、轮廓特征等）进行多维数据整合与协同分析，可以有效挖掘原始图像中相关特征信息、增强特征信息的可信度，由于其是基于特征的方法，所以其较之像素级图像融合方法具有更高的鲁棒性和自适应性。刘峰[43]等人针对复杂环境下海上舰船目标单波段特征识别能力不足的问题，利用可见光、中波红外和长波红外特征图像融合技术提出一种基于区域协方差矩阵的多波段特征融合方法，分别为可见光图像和中波、长波红外图像设计不同维度的特征向量，同时通过协方差矩阵将多个特征融合保证不同目标的差异性。Hui Li 等人提出 DenseFuse 模型[44]，其编码网络可以确保输入层精准连接到输出层，在图像的融合过程中可以获得更多层次的特征信息，提高目标检测的准确性。之后 Hui Li 等人继续革故鼎新，提出基于嵌套连接网络和空间通道注意力的 NestFuse 模型[45]及基于残差结构的残差融合网络 RFN-Nest[46]。其中 RFN-Nest 在嵌套连接的基础上，通过细节保持损失函数和特征增强损失函数训练 RFN，实现图像融合端到端的网络，RFN-Nest 网络结构图如图 1-3 所示。

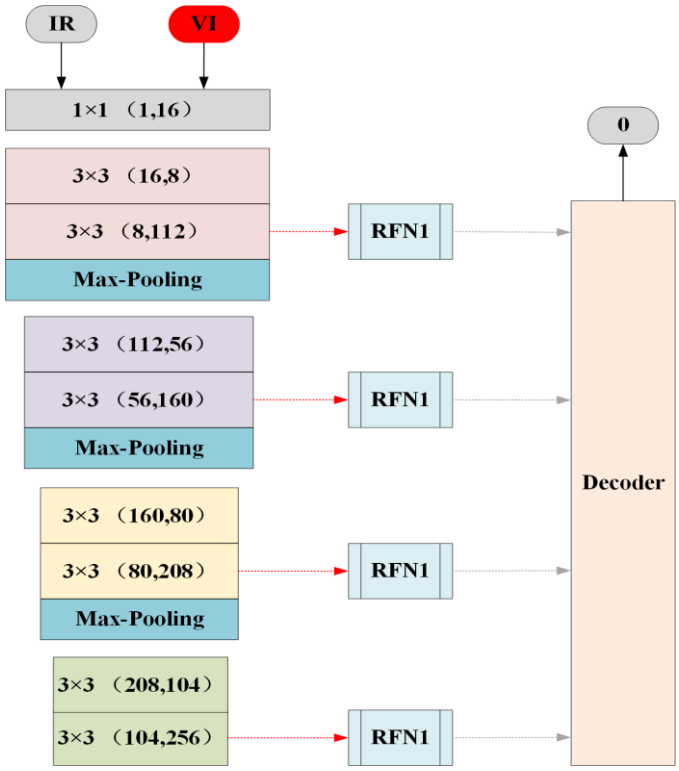


图 1-3 RFN-Nest 网络结构图

1.3 主要研究内容及章节安排

1.3.1 主要研究内容

与常见可见光条件下的行人检测不同，基于夜间场景下的行人检测往往面临的是夜间道路无可见光和能见度更低的场景，例如夜间完全无照明道路场景、夜间昏暗道路场景，这一类场景通常具有可见光较弱甚至无可见光的特点，传统情况下依靠可见光充足道路场景的行人检测方法难以有效的检测到行人目标。

本文主要围绕基于夜间场景下红外图像行人检测和红外-可见光图像融合行人检测展开研究。通过查阅近些年国内外资料并分析当前行人检测可能存在的问题，进行相关对比实验筛选出更适合夜间道路场景行人检测的 YOLOv8 算法，着重对 YOLOv8 目前面临的夜间道路场景下的红外图像行人检测存在网络模型较大、检测精度降低的问题提出相应的解决方案，对于夜间道路场景中可能存在部分可见光的情况下，如何将红外图像的优势与可见光图像的优势相加后再进行相应的行人检测。本文研究内容主要可以划分为以下三个方面：

（1）筛选出道路驾驶场景较为全面的 FLIR 红外目标检测数据集中部分夜间红外图像数据和具有一定可视条件下的红外-可见光图像，分别划分为本文实验所需的两个数据集，并人工制作标签。对预建立的红外数据集进行预处理，使用 DDE 增强算法进行预处理，能更好地平衡图像整体的灰度，抑制过亮和过暗区域对全局显示的影响，增强了低对比度区域的景物细节。

（2）针对当前红外图像行人识别容错率较低、网络模型较大的问题，提出一种基于 YOLOv8 网络模型改进的轻量级红外图像行人检测算法 RT_YOLOv8。首先采用 RepGhostNet 作为主干网络，提升模型在总体上对行人的特征提取能力；引入 Triplet 注意力机制，通过宽度(W)、高度(H)、通道(C)三维度的交互，使得网络更好的提取目标特征信息；使用可变形卷积 DCNv2 重新构建 C2f 模块来提升网络的检测能力；在 Detect 模块中引入 PConv 替换部分常规卷积，减少冗余计算和内存访问，有效地提取空间特征。

（3）针对具有一定可视条件的夜间道路场景情况下，采用红外-可见光图像中期融合后行人检测。本文基于红外-可见光图像的融合设计了一种夜间场景下的跨模态双分支行人检测网络。通过在 SPPF 模块中引入 LSKA 来提高对跨尺度特征的感知能力；通过消融实验的结果，引入第四章提出的 C2f 模块、Triplet 注意力机制和 Detect 模块将其融合于跨模态行人检测网络中；最后针对基准网络损失函数存在的问题，将其替换为 MPDIoU 损失函数，加快模型的收敛速度。

1.3.2 章节安排

本文主要针对夜间道路场景行人检测的研究，针对现有红外图像行人检测与红外-可见光图像跨模态行人检测的研究不足，针对夜间智能驾驶系统的技术难点，构建了层次化研究体系，本文共设六章递进式研究，每一章的具体研究内容如下：

第 1 章 绪论。本章系统地阐述研究背景及在智能驾驶系统中的关键作用及解析当前行人检测技术的发展历程及趋势，重点阐述基于红外图像的行人检测与基于红外-可见光跨模态的行人检测领域的研究进展，同时明确本文的核心研究框架与章节逻辑体系。

第 2 章 目标检测相关理论概述。本章系统地阐述有关目标检测的深度学习理论知识，然后介绍了基于深度学习的目标检测部分经典网络模型，最后介绍了目标检测相关评价指标。

第 3 章 图像预处理及数据集的构建。首先介绍红外成像技术的相关理论知识和当前主流法红外行人检测开源数据集（如 FLIR、KAIST 多光谱数据集）。然后基于开源红外图像数据集，建立适合本文夜间道路场景的行人检测数据集及相关数据集前期预处理，其中包括：红外图像行人数据与红外-可见光图像行人数据的筛选。红外图像数据的前期增强处理、相关对比实验验证增强算法的有效性、人工标注行人标签文件，最后划分为红外图像行人检测数据集及红外-可见光图像行人检测数据集。

第 4 章 基于红外图像的行人检测算法研究。本章主要介绍了在 YOLOv8 模型的基础上，对其网络模型进行优化，从而达到兼顾行人检测实时性与准确性的实际需求。为了保证行人检测准确性的需求，采用 RepGhostNet 作为主干网络，提升模型在总体上对行人的特征提取能力，并且在主干网络末端引入 Triplet 注意力机制，通过多维度的交互，使得网络更好的提取目标特征信息，最后使用可变形卷积 DCNv2 重新构建 C2f 模块来提升网络的检测能力；为了保证行人检测实时性的要求，在 Detect 模块中引入 PConv 替换部分常规卷积，减少冗余计算和内存访问；最后进行相关的消融实验及行人检测算法的对比实验。

第 5 章 基于红外-可见光图像跨模态融合的行人检测算法研究。本章提出一种基于中期融合的跨模态行人检测网络，通过双分支特征提取的方式分别提取红外图像和可见光图像的特征后经过中期融合将提取到的特征进行融合；通过消融实验的验证引入 RT_YOLOv8 中的 C2f 模块、Triplet 注意力机制和 Detect 模块，将其融合于跨模态网络中；并且在特征提取阶段将 LSKA 融入到 SPPF 模块中，增强对跨尺度特征的感知能力；然后通过对损失函数进行改进，引入 MPDIoU 损失函数，加快模型收敛速度；最后通过消融实验和对比实验来验证算法的有效性。

第 6 章 总结与展望。文章系统地梳理了本文研究内容的创新成果。客观分析当前夜间行人检测存在的局限性与研究过程中存在的技术难点，在此基础上提出后续研究的改进路径与展望。

第2章 目标检测相关理论概述

本章系统阐述研究涉及的目标检测核心理论及方法。为后续研究奠定技术基础与理论支撑。首先介绍卷积神经网络的相关概念，主要包括：神经网络、激活函数、损失函数以及常见的目标检测神经网络结构，最后介绍本文实验所采用的评价指标。

2.1 深度学习理论知识

2.1.1 卷积神经网络

（1）神经网络概述

神经网络是一种通过模拟人脑的神经网络从而实现人工智能的机器学习技术。人工神经网络通过借鉴生物神经网络的拓扑结构，由高密度节点及层级化连接构成，通过重参数互联构建分布式计算架构，对其连接处设置的权重值来达到数据处理的目的。所设置的权重值可以通过后续神经网络的不断训练过程进一步优化，从而达到最优效果。其中卷积神经网络[47-49]通常包括卷积层、池化层、全连接层、激活函数、损失函数等。神经网络的基本结构如图 2-1 所示。

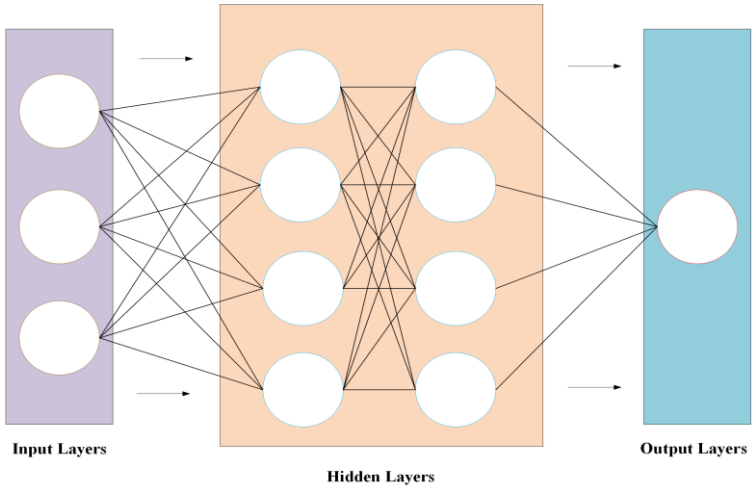


图 2-1 神经网络基本结构图

通常将神经网络分为三部分，其中输入层是神经网络处理数据的开端，其作用为用来接收数据并且将其传递至隐藏层；隐藏层是位于输入层与输出层的中间

部分，其作用为将输入数据转换为更高层次的特征。在卷积神经网络中，卷积层和池化层共同组成了隐藏层，在循环神经网络中，又由循环层构成。神经网络的处理问题能力强弱跟隐藏层的神经元数量密不可分，事物皆有利有弊，在神经网络处理问题能力提高的同时带来的问题则是网络模型会随之变大，大大增加了网络的计算复杂度；输出层是神经网络的最后一层，输出结果代表了神经网络对输入数据的预测结果。

（2）卷积层

在卷积神经网络模型中，卷积运算是将数据窗口与图像中逐个元素相乘后求和的过程，数据窗口为一组固定的权重值，可以看成一个特定的过滤器(Filter)。如图 2-2（a）所示，左侧为输入矩阵，中间为过滤器或卷积核(Convolutional Kernel)。卷积核通过固定的步长(Stride)在输入矩阵中移动，分别进行相乘后相加的运算，得到右侧的输出矩阵。输入矩阵为 3×3，在输入矩阵中使用 2×2 卷积核进行卷积操作，步长设置为 1，零填充(Zero Padding)，最后得到一组 2×2 的矩阵，即 0×1+1×2+2×4+3×5=25。输出矩阵的其他数值计算方式如图 2-2（b）、（c）、（d）所示。最后这些输出矩阵组成一个新的特征图(Feature Map)，卷积层的输出特征计算公式如式(2-1)所示。

$$Y_{i,j}^{k} = f\left(\sum_{l=1}^{C}\sum_{m=1}^{K}\sum_{n=1}^{K} W_{m,n,l}^{k} X_{(i-1)s+m,(j-1)s+n,l} + b_k\right) \tag{2-1}$$

其中，X、Y 分别为卷积层输入、输出特征图，K 为卷积核的大小，s 为卷积核的 Stride，W 为卷积核的权重，b 为卷积核的偏置项。

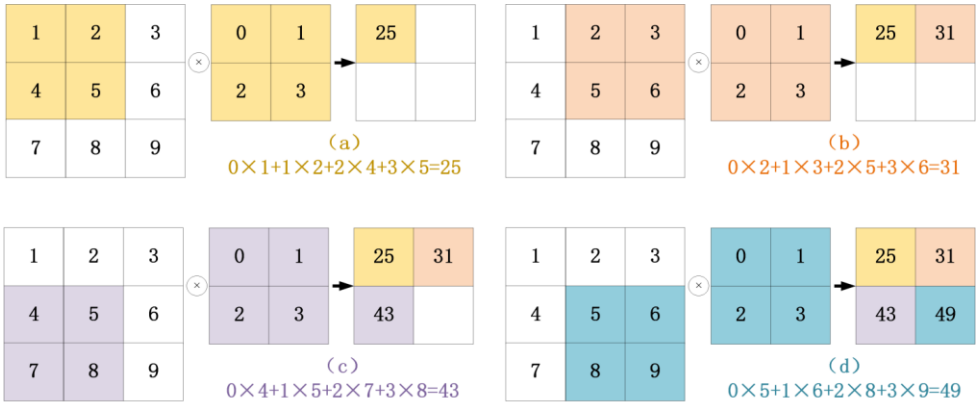


图 2-2 卷积示意图

卷积层在神经网络中扮演的角色为提取输入数据的特征。在此过程中可以捕

捉到输入数据的局部模式和空间不变性，例如边缘、纹理和形状等特征，减少了对于输入数据位置的敏感性。在整个神经网络中卷积层带来的优势为：减少参数量，避免后续的过拟合；局部连接，由于每个卷积核只对应输入矩阵的小部分区域，所以提高了数据处理效率。

（3）池化层

神经网络的池化层(Pooling Layer)主要目的是用来降低数据的空间大小，减少由于数据处理中产生的不必要计算量，并且在一定程度上有效的避免了过拟合的发生。通过在神经网络中对卷积层输入的预处理数据进行维度降低处理，在最大程度保证输入数据的有效信息的同时减少冗余计算量。

池化层分为平均池化(Average Pooling)与最大池化(Max pooling)，两种数据处理的区别在于平均池化将输入矩阵中的数值取平均值后输出，最大池化是将输入矩阵中的数值取最大值后输出。平均池化与最大池化过程如图 2-3 所示。

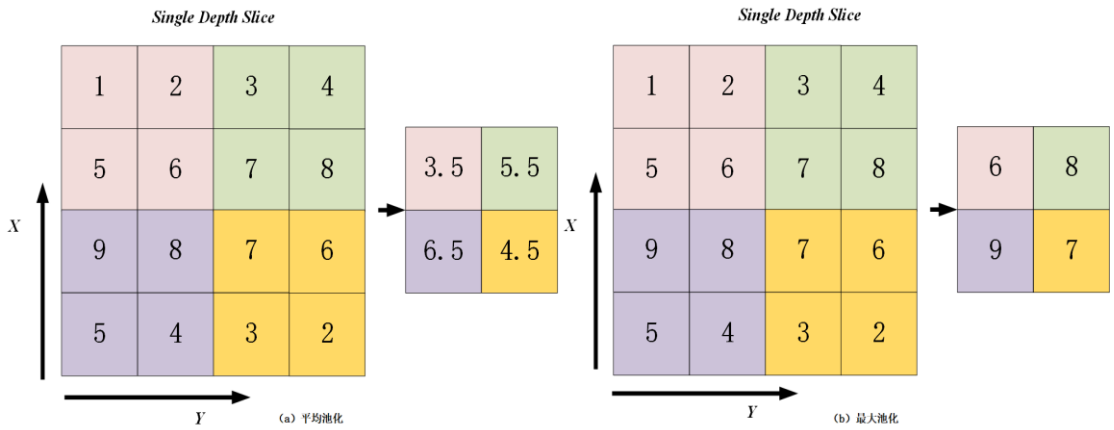


图 2-3 池化计算过程

（4）全连接层

神经网络中的全连接层(Fully Connected Layer, FC)充当了一个分类器的角色。卷积层与池化层通过多层次特征提取机制将未处理的输入信息转换至隐藏特征空间，而全连接层则承担特征整合与抽象表征的关键作用。在全连接层里面，前序卷积层所提取的局部特征会与后续神经元建立全连接关系，通过权重矩阵的线性组合与激活函数的非线性变换，最终形成具有高层次语义信息的一维特征向量。生成后的特征向量会与所设定的权重值进行数学运算，最后将输出参数输入到分类器中。

如图 2-4 所示，假设神经网络模型已经训练完成，全连接层通过对输入图片进行分类，最后基于分类的特征可以清楚知道输入图片的类型。

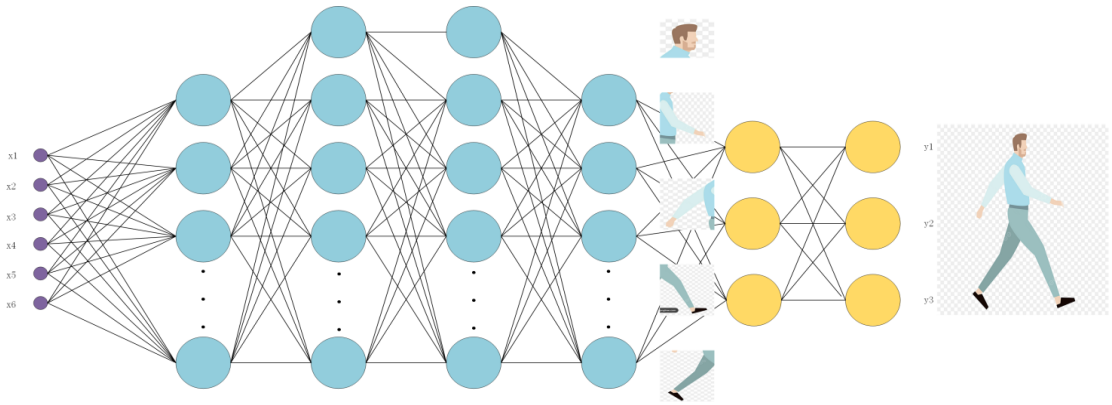


图 2-4 全连接层示意图

2.1.2 激活函数

激活函数[50]在神经网络中承担非线性变换的关键工作，在每个神经元之间通过引入非线性后将其输出映射到指定的范围内。如果没有激活函数的参与，每一层的输出都是由上层至下层的线性函数，这样输出函数永远是输入的线性变换组合，称为感知机(Perceptron)。从神经元通过激活函数得到最终的输出示意图如图 2-5 所示。

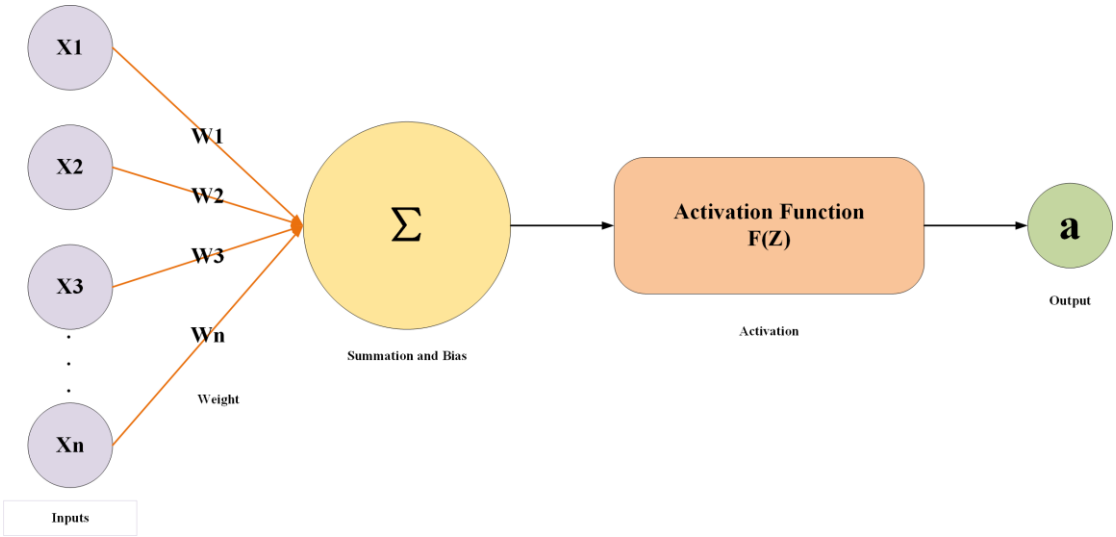


图 2-5 激活函数示意图

激活函数可以分为两大类：饱和激活型函数、非饱和型激活函数。

从数学定义角度分析，对于给定激活函数 $f(x)$，当自变量 n 无限趋近于正无穷时其导数趋近于 0，则该函数满足右饱和激活函数特性。当 n 无限趋近于负无穷时其的导数趋近于 0，则该函数满足左饱和激活函数特性。仅当同时满足双侧饱和条件时，方可定义为严格意义上的饱和激活函数。典型代表如 Sigmoid 函数与双曲正切函数(Hyperbolic Tangent Function, Tanh)等。其中 Sigmoid 激活函数[51]的数学表达式如式(2-2)所示，其函数曲线呈显著 S 型分布特征，如图 2-6（a）所示。由图可知 Sigmoid 函数为一种 S 型函数，值域为(0, 1)，首尾两端无限趋于 0 和 1。这种非线性响应机制使得该函数在输入信号处于极值区域时呈现显著的特征差异。从数学定义角度来看，函数属于连续可导，其导数无限趋于 0，这时候会导致梯度消失的情况出现。

Tanh 激活函数[52]的数学表达式如式(2-3)所示，Tanh 激活函数图像如图 2-6（b）所示。由图可知其函数图形为双正切曲线，其值域为(-1, 1)，且经过坐标原点，此时会得到更快的收敛速度。然而其导数仍然是无限趋于 0，同样也存在梯度消失的情况。

$$Sigmoid(x)=\sigma(x)=\frac{1}{1+\exp(-x)} \tag{2-2}$$

$$Tanh(x)=\frac{\exp(x)-\exp(-x)}{\exp(x)+\exp(-x)} \tag{2-3}$$

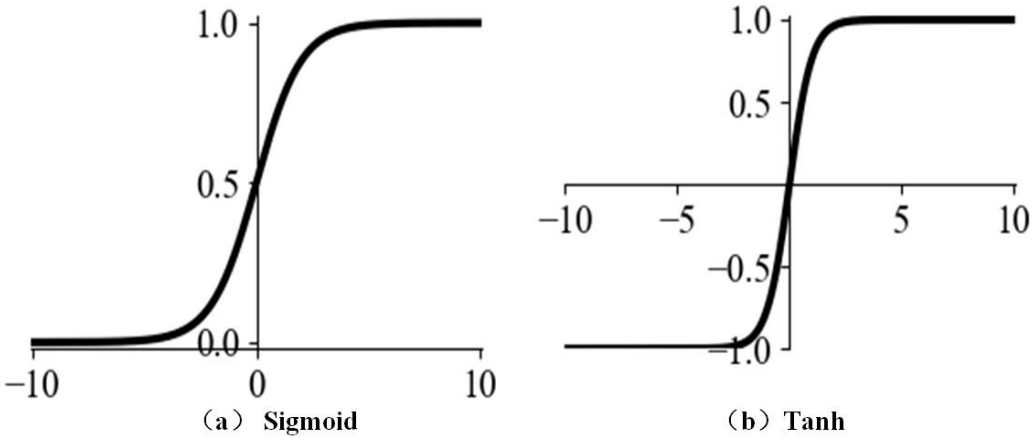


图 2-6 Sigmoid 和 Tanh 激活函数

反之，在激活函数 $f(x)$ 即不满足右饱和也不满足左饱和的情况称为非饱和函数。相较于饱和激活函数，非饱和激活函数可以有效解决由于神经网络堆积的

层数较多的情况下带来的梯度消失等问题，并且能够有效提升收敛速度。常见的非饱和激活函数有 ReLU(Rectified Linear Unit)、PReLU(Parametric Rectified Linear Unit)等。ReLU 激活函数[53]的数学表达式如式(2-4)所示，ReLU 激活函数图像如图 2-7（a）所示。其值域为(0, +∞)。从函数图像上来看，函数在(0, +∞)时，此时输出值等于输入值，然而在(-∞, 0)时，无论输入值如何变换输出值依然为 0。并且在(0, +∞)区间内的情况下，导数为 1。有效解决了饱和函数未解决的梯度消失问题。具有较快的收敛速度，适合于更复杂的神经网络。

PReLU 激活函数的数学表达式如式(2-5)所示，PReLU 激活函数[54]通过引入可学习参数 a 的变换量来调节不同神经元之间的梯度，通过设置 num_parameters 模块来控制参数 a。PReLU 激活函数图像如图 2-7（b）所示。其值域为(－∞, +∞)，仍然在(0, +∞)时，输出值等于输入值，保留负激活数据。

$$Re\,LU(x)=(x)^{+}=Max(0,x) \tag{2-4}$$

$$P\,Re\,LU(x)=\begin{cases}x, x\geq 0\\ ax, otherwise\end{cases} \tag{2-5}$$

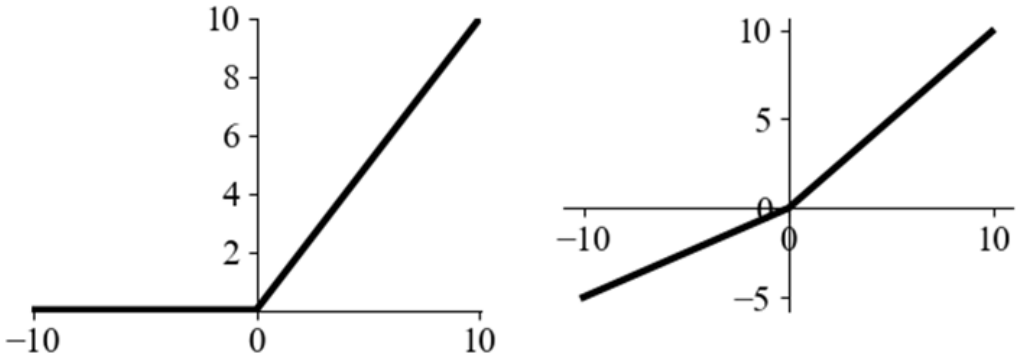


图 2-7 ReLU 和 PReLU 激活函数

2.2 经典目标检测算法介绍

2.2.1 YOLOv5

YOLOv5 作者 Glenn Jocher 在 YOLOv4 的基础上进一步改进，具有四个版本，然而四个版本的基本结构一样，区别是参数 depth_multiple 和 width_multiple

不同，其中 YOLOv5s 的 depth_multiple 和特征图的 width_multiple 最小，其他三种在此基础上不断深入。YOLOv5 架构在继承前代框架基础上实现了多维度技术升级，其核心创新点体现在：

（1）引入 Mosaic 四图融合增强技术，提升小目标训练样本密度（数据多样性+230%）。集成动态优化策略：自适应锚框聚类算法（k-means++优化）和智能图像缩放技术（填充损耗降低 40%）；

（2）Backbone 端使用了 YOLOv2 的 Passthrough 模块命名为 Focus，与 YOLOv4 不同的是，YOLOv4 只有 Backbone 端使用了 CSP(Cross Stage Partial)网络模型，而 YOLOv5 设计了两套 CSP(CSP1_X、CSP2_X)网络模型，分别运用在主干网络和颈部网络中，CSP 网络模型结构如图 2-8 所示；

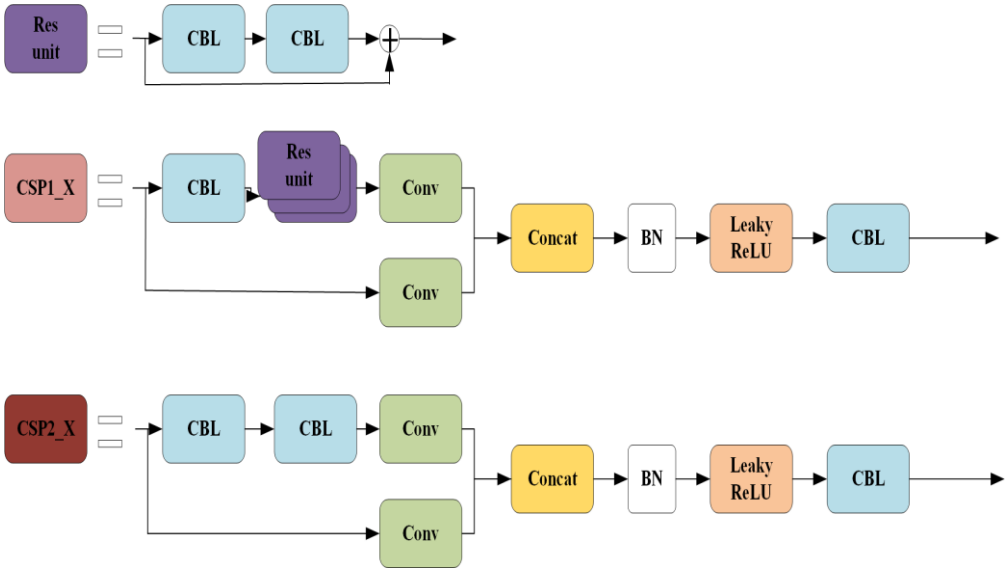


图 2-8 CSP 网络模型结构图

（3）Neck 端与 YOLOv4 同样采用 FPN+PAN 网络模型，采用 CSP2_ResBlock 模块重构特征金字塔，相比 YOLOv4 的常规卷积组合，计算复杂度降低 18%(FLOPs: 106.3→87.2)，特征融合路径优化为双向跨尺度连接，mAP50 提升 2.6%；YOLOv5 网络结构图如图 2-9 所示。

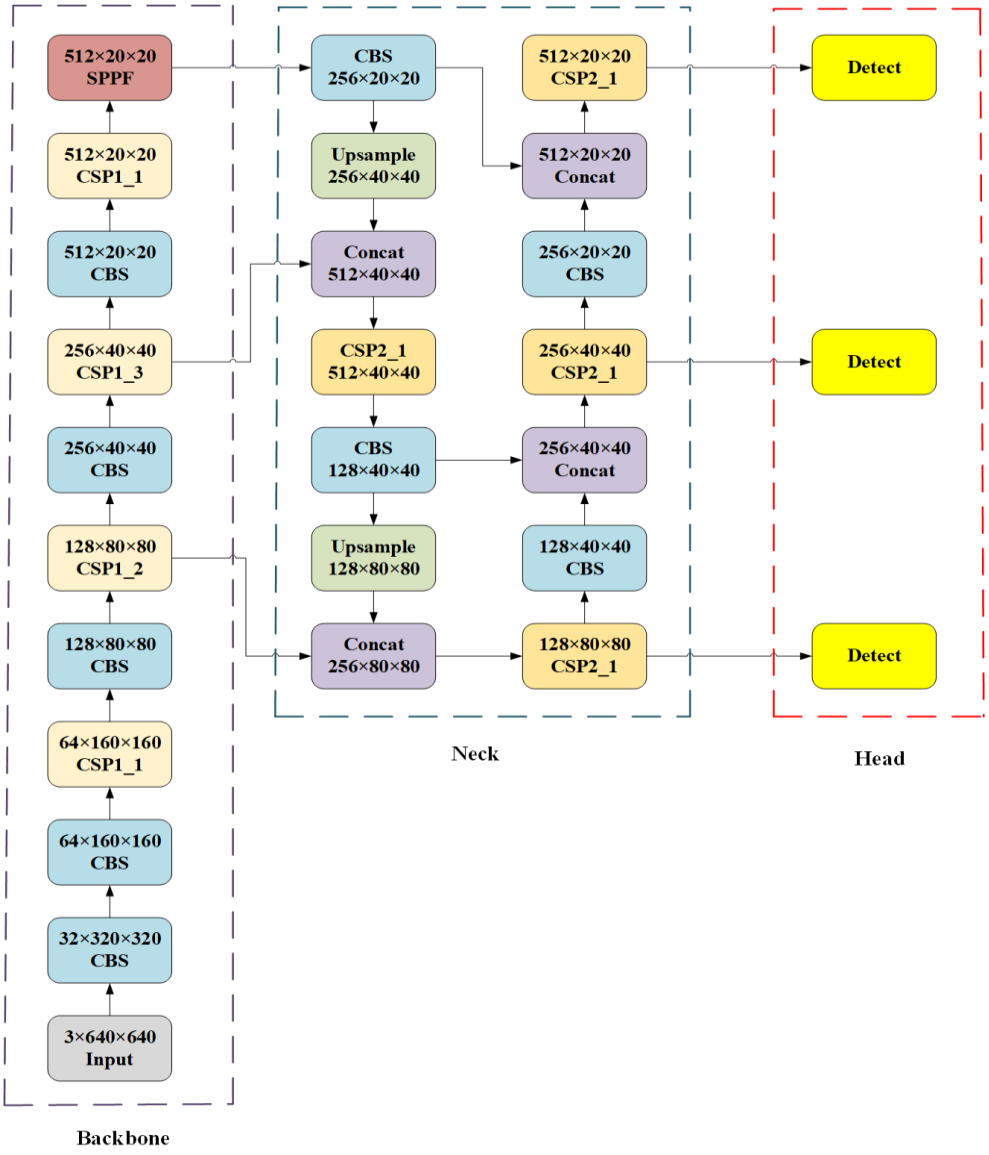


图 2-9 YOLOv5 网络结构

（4）Head 端使用 CIoU_LOSS 作为 Bounding Box 的损失函数，CIOU 在 DIOU 的基础上增加了高宽比例，CIoU_LOSS 计算公式如式(2-6)、(2-7)、(2-8)、(2-9)所示：

$$L_{CIoU}=1-IoU+\frac{\rho^2\left(B_{gt},B_{pred}\right)}{C^2}+\alpha v \tag{2-6}$$

$$IoU=\frac{B_{gt}\cap B_{pred}}{B_{gt}\cup B_{pred}} \tag{2-7}$$

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}}-\arctan\frac{w_{pred}}{h_{pred}}\right)^2 \tag{2-8}$$

$$\alpha = \frac{\nu}{(1-IoU)+\nu} \tag{2-9}$$

其中，IoU 为预测框与标注框的交并比，B_{gt} 为真实框，B_{pred} 为预测框；$\rho^2(B_{gt}, B_{pred})$为预测框到真实框中心点的欧氏距离；α 为平面参数；ν用来衡量长宽比是否一致。

（5）YOLOv5 采用 *GIoU*(Generalized Intersection over Union)作为边界框回归的损失函数。*GIoU* 及 *GIoU_Loss* 的计算公式如式(2-10)、(2-11)所示：

$$GIoU = IoU - \frac{|C \setminus (A \cup B)|}{|C|} \tag{2-10}$$

$$GIoU\_loss = 1 - GIoU \tag{2-11}$$

其中，C 为一个包含 A 框和 B 框的最小框，$C \setminus A \cup B$ 表示 C 面积减去 $A \cup B$ 的面积。

在 YOLOv5 中，分别由 Bounding Box 回归损失、目标置信度损失、类别损失构成，整体的损失函数如式(2-12)所示：

$$L_{v5}(t_p, t_{gt}) = \sum_{k=0}^{K}[\alpha_k^{balance}\alpha_{box}\sum_{i=0}^{s^2}\sum_{j=0}^{B}II_{kij}^{obj}L_{CIOU} + \alpha_{obj}\sum_{i=0}^{s^2}\sum_{j=0}^{B}II_{kij}^{obj}L_{obj} + \alpha_{cls}\sum_{i=0}^{s^2}\sum_{j=0}^{B}II_{kij}^{obj}L_{cls}] \tag{2-12}$$

其中，K，s^2，B 分别为输出特征图、cell、每个 cell 上 anchor 的数量；II_{kij}^{obj} 表示第 k 个输出特征图；t_p, t_{gt} 分别表示预测向量、Ground_Truth 向量；$\alpha_k^{balance}$ 表示平衡尺度的输出特征图权重值。

2.2.2 YOLOv6

YOLOv6 由美团视觉智能部开发的目标检测网络，为了进一步提升目标检测的性能使用了深度可分离卷积(Depthwise Separable Convolution, DSConv)等技术。

YOLOv6 对主干网络和颈部网络都进行了重新设计，头部网络则在 YOLO_X 中的 Decoupled Head 的基础上进行优化，从颈部网络输出三个分支后对输出 feature map 通过 BConv 进行特征融合，BConv 由 Conv、BN 和 SiLU 组成，然后分支一通过 BConv 和 Conv 完成分类任务，分支二通过 BConv 后分别再通过 Conv 进行边框回归和通过 Conv 进行前后背景的分类，最后输出最终预测结果。YOLOv6 在 YOLOv5 的基础上引入 RepVGG 结构重参模块，由于之前的 YOLO

系列对 GPU 硬件设施利用率较低，然而 RepVGG 却可以很好地解决这个问题，具有内存带宽、编译优化等特性。RepVGG 训练过程中利用多分支结构为 YOLOv6 带来更高的检测精度。YOLOv6 网络结构如图 2-10 所示。

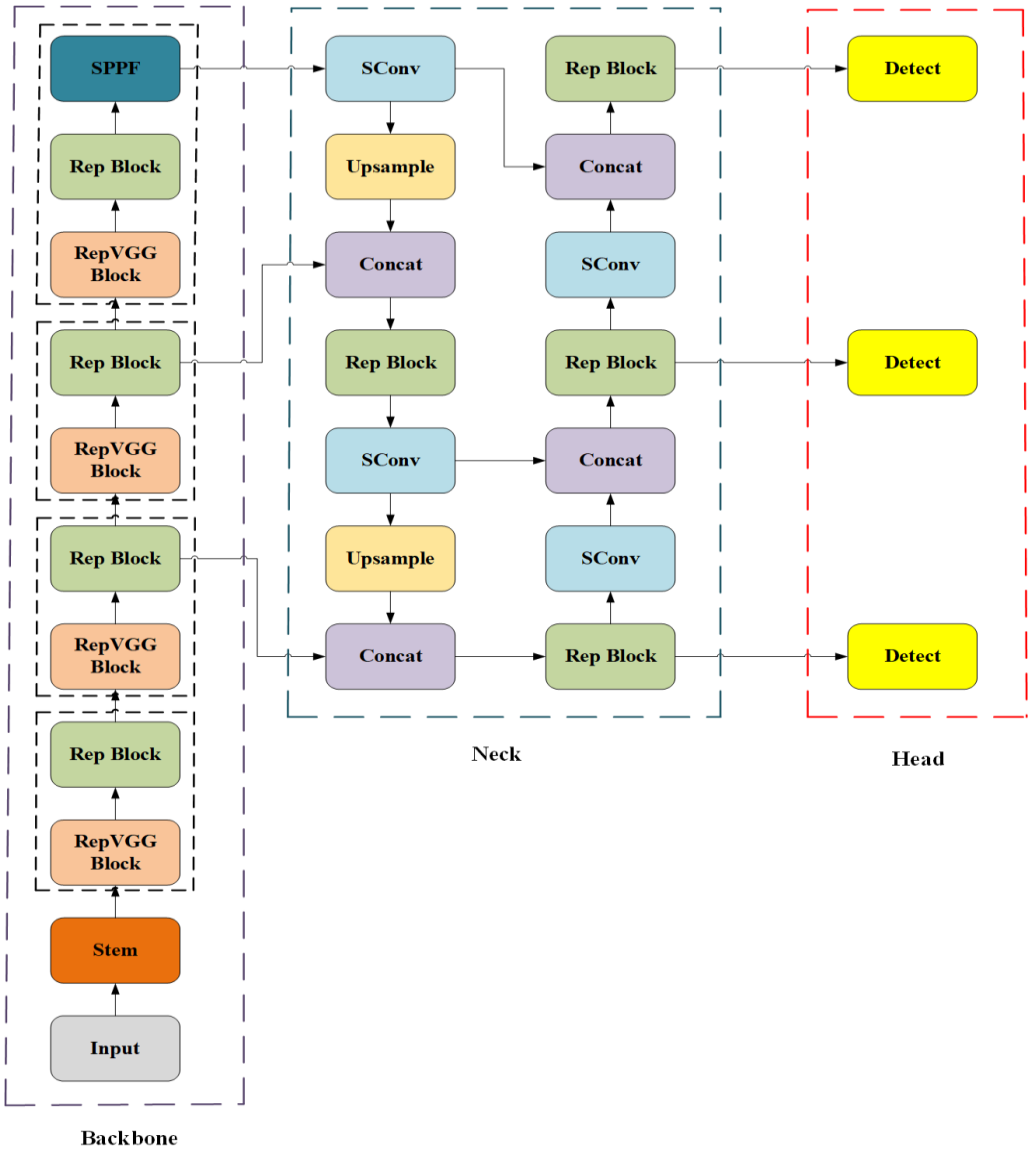


图 2-10 YOLOv6 网络结构

2.2.3 YOLOv7

YOLOv7 作者和 YOLOv5 为同一人，YOLOv7 仍然基于 anchor based，同时增加了 E-ELAN 结构，在头部模块增加了 Aux_detect，通过对预测结果的进一步筛选，增加模型整体检测精度，类似双阶段(two stage)的检测方法。

其主要改进点为将模型重参数化引入到模型网络架构中，主要对 Conv+BN

层及其他卷积进行特征融合，最后成为一个完整的卷积模块，其 Conv+BN 层融合的计算公式如式(2-13)所示。

$$\begin{aligned} fuse\_conv\_bn &= \frac{\gamma((wx+b)-m)}{\nu}+\beta \\ &= \frac{\gamma(wx+(b-m))}{\nu}+\beta \\ &= \frac{\gamma w}{\nu}x+\frac{(b-m)}{\nu}\gamma+\beta \end{aligned} \tag{2-13}$$

其中 w 和 b 分别代表权重和参数，为 Conv 参数；m、ν 分别代表输入均值、输入标准差；γ、β 为可学习的参数值。

在 YOLOv7 中检测时将头部模块浅层特征提取出来作为 Aux_detect，深层特征则为网络的最终输出 Lead Head；其标签分类方式为 YOLOv5 的跨网格搜索并且用到了 YOLOx 的匹配方法；YOLOv7 在 5FPS~169FPS 的范围内，超过之前的目标检测算法。YOLOv6 网络结构图如图 2-11 所示。

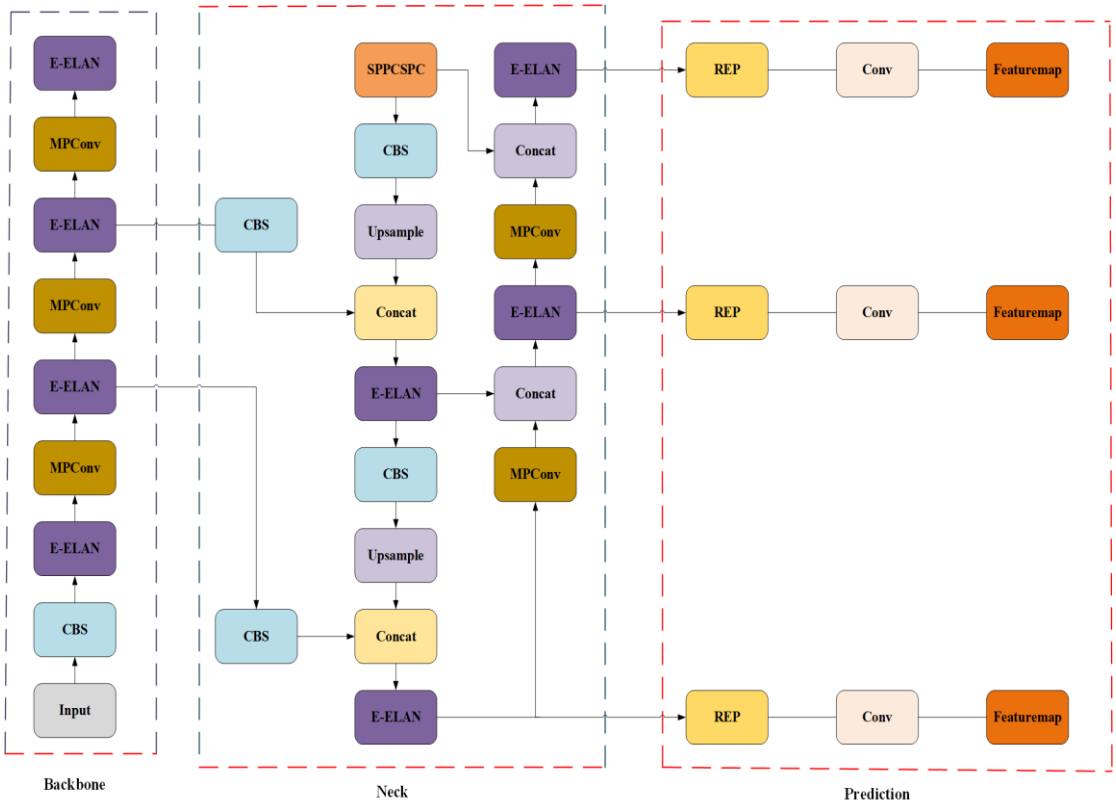


图 2-11 YOLOv7 网络结构

2.3 评价指标

为了验证算法的性能。实验采用精确率(Precision)、召回率(Recall)、平均精度(Mean Average Precision, mAP)、模型参数量(Parameters)和浮点型计算量(FLOPs)为评价指标。

精确度(Precision)代表模型预测中的正确率，*TP* 是模型预测的真正例，*FP* 是模型预测的假正例。精确度表达式如式(2-14)所示：

$$Precision = \frac{TP}{TP + FP} \tag{2-14}$$

召回率(Recall)代表模型真正例样本数与真实正例总数的比值，*FN* 为模型预测的假负例。召回率表达式如式(2-15)所示：

$$Recall = \frac{TP}{TP + FN} \tag{2-15}$$

平均精度(Average Precision, AP)代表模型对于这类目标检测的准确度，为了兼顾精确度和召回率通常需要借助 PR 曲线来分析，以 Precision 为 X 轴、Recall 为 Y 轴。通过曲线求积分来计算面积，得出的面积结果则为 *AP* 值，$p(r)$ 为某个类别的平均精度。总类别平均精度的表达式如式(2-16)所示：

$$AP = \int_0^1 p(r)dr \tag{2-16}$$

平均精度均值(mAP)指所有类别的平均 *AP* 值，在计算出所有类别的 *AP* 后除以类别总数，得到的各类别 *AP* 的平均值则为 mAP。平均精度均值表达式如式(2-17)所示：

$$mAP = \frac{1}{k}\sum_{i=1}^{k} AP_i \tag{2-17}$$

模型参数量(Parameters)指所有需要通过学习（训练）确定的变量总数，直接影响模型的表达能力、计算资源消耗和模泛化能力。

浮点型计算量(FLOPs)是衡量模型计算复杂度的关键指标，表示模型完成一次前向传播所需的浮点计算次数(Floating Point Operations)。是评估模型运行效率、硬件资源需求和实际部署可行性的重要依据。

2.4 本章小结

本章围绕深度学习目标检测技术展开系统性阐述，逻辑结构分为三个递进层次：首先，深入解析深度学习的核心组件及原理，包括卷积层、池化层、全连接层，以及激活函数（如 ReLU、Sigmoid）的非线性建模作用，特别强调这些组件在特征提取中的协同工作机制。其次，系统梳理部分 YOLO 系列算法，着重分析各代算法在骨干网络设计、损失函数优化、数据增强策略等方面的突破。最后，构建目标检测评价体系，涵盖基础指标精确率、召回率和平均精度等，形成兼顾准确率与实时性的立体化评估维度。该理论框架为后续的模型优化、算法对比研究奠定了坚实的理论基础与方法论指导。

第3章 图像预处理及数据集的建立

3.1 引言

目前开源目标检测数据集有 KAIST、FLIR、LLVIP 等，然而其本文研究内容是基于道路场景的夜间行人检测，所以可供实验直接使用的场景。所以本文基于 FLIR 数据集的基础上，从中筛选出部分夜间道路场景的红外图像和部分与之对应的可见光图像作为本文实验数据集，并且自动划分训练集和验证集。本文所涉及到的红外图像预处理是指在使用其他算法分析之前对部分红外图像进行特征增强等操作。对红外图像采集系统存在的硬件限制与环境变量干扰，需建立定向增强处理机制来优化目标特征表征。具体而言，由于探测器噪声、大气衰减等客观因素造成图像信噪比(SNR<35dB)显著降低，导致有效特征提取存在障碍。因此本章采用 DDE 特征增强算法对部分红外图像进行特征增强，经过该预处理之后能够有效增强行人特征，并且通过常见评价指标对处理图像进行验证，从而证明算法的有效性。

3.2 红外成像技术理论

3.2.1 红外成像原理

由于人的眼睛可以捕捉的可见光谱(0.4-0.8um)较窄，所以自然界中物体的光谱只要不在此范围内，凭借肉眼是无法捕捉到的。根据现代物理学的原理，凡是温度高于-273℃的任何物体都在时刻向外界环境释放电磁波。本文数据集中红外图像通过 FLIR 公司的红外热成像仪采集而来。

红外热成像仪[55]将目标或者目标的不同部位通过反射或者辐射的红外热能量的差异情况通过二维图像呈现出来或者记录下来。按照其工作原理通常分为制冷型光机扫描热像仪、非制冷焦平面热像仪。

制冷型光机扫描热像仪工作原理：通过热像仪镜头捕捉到待测目标自身发出的电磁波后经过水平扫描棱镜、垂直扫描棱镜聚光和滤光将电磁波聚集到红外探

测器上，后通过信号处理系统放大为电信号输出到图像生成器。非制冷焦平面热像仪工作原理：由于仪器本身红外探测器为二维平面形状且具备电子扫描功能，待测目标只需要简单的聚焦成像即可在图像生成器上出现，其原理类似于相机。

红外热成像仪通常由光学系统、探测器、制冷系统、信号处理系统、输出接口、图像生成器组成，红外热成像系统的结构如图 3-1 所示。

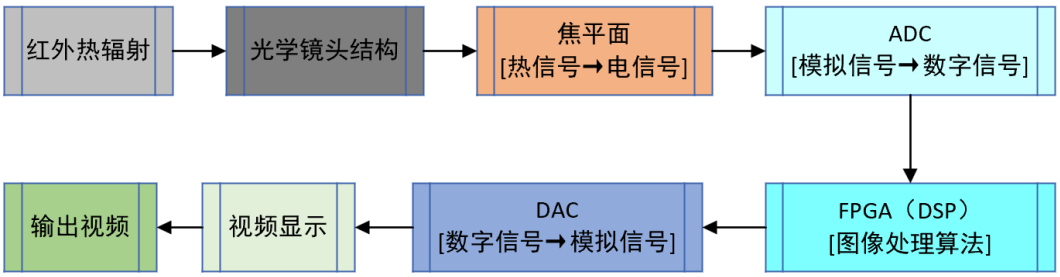


图 3-1 红外热成像系统结构

3.2.2 红外图像特点

红外图像[56]灰度信息与可见光相差较大，由于物体对不同波长的辐射的反射、透射、发射情况不同，同时红外图像灰度信息还受环境等不可控因素影响。灰度直方图本质上是表征数字图像灰度分布的统计工具，其核心功能是呈现各灰度级在图像中的分布频次，直观地反映出图像灰度概率分布特性。该灰度图生成统计图的坐标系构成具有明确的物理意义：横坐标表征 0-255 区间内的离散灰度级，纵坐标对应各灰度级在图像中归一化出现频率次数。若以数学表达式描述，假设一个图像的灰度值用 X_i 表示，每个灰度值的数量用 N_{X_i} 表示，其概率密度分布函数 $P(X_i)$ 公式如式(3-1)所示。

$$P(X_i)=\frac{N_{X_i}}{\sum_{i=0}^{255}N_{X_i}} \tag{3-1}$$

与可见光图像相比较，红外图像灰度直方图具有一定的规律性。如图 3-2 所示，左图是该图的红外图像，右图是其对应的灰度直方图，能够看出该图整体较为灰暗，部分细节及较远处小目标不能很好地呈现处理。如图 3-3 所示，左图是可见光的图像，右图是其对应的灰度直方图。

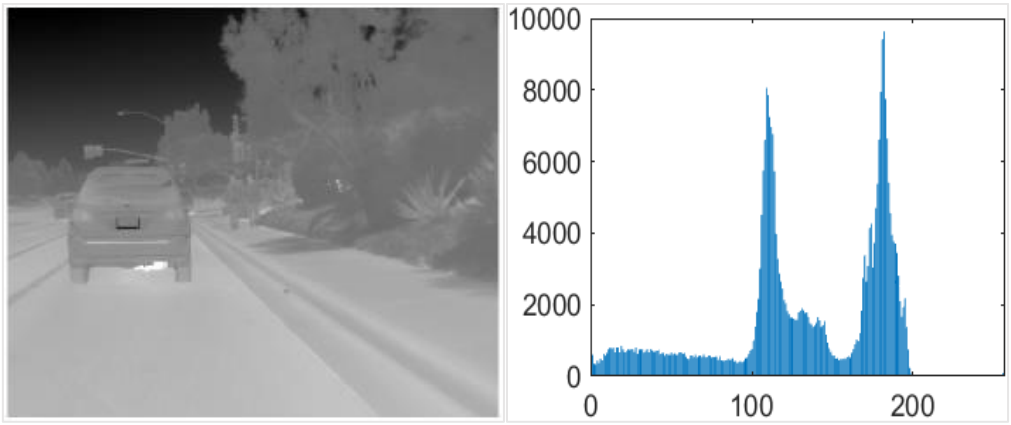


图 3-2 红外图像与其对应的灰度直方图

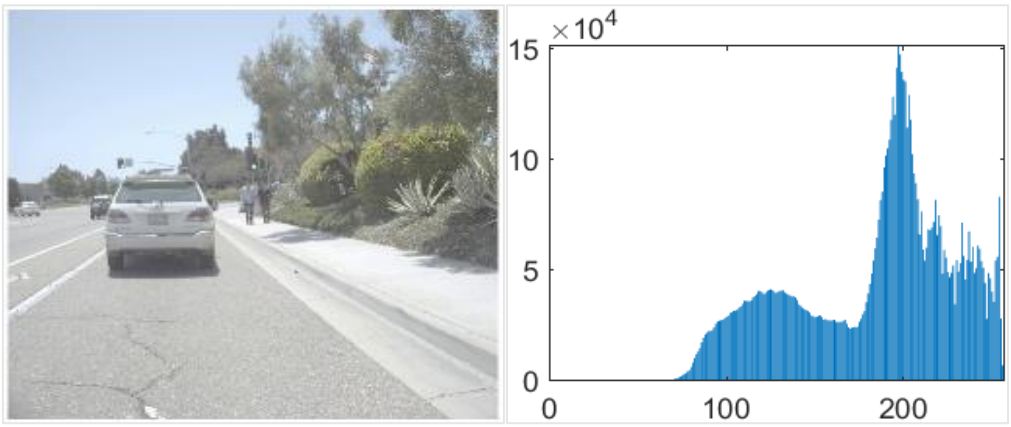


图 3-3 可见光图像与其对应的灰度直方图

通过对比两种图像的灰度直方图可以发现，红外图像呈现出明显的灰度动态范围压缩特征，其像素灰度值主要集中分布在较为狭窄的区间内，仅覆盖全图灰度谱的有限部分。与之形成对比的是，可见光图像的灰度级分布展现出更好的均衡化特征，其灰度值分布于整个可呈现范围内，占据了灰度谱的主题区域。此外还可以发现红外图像普遍呈现显著的双峰分布形态，这与该类型图像灰度级聚集压缩的特性直接相关；而可见光图像由于灰度分布离散度较高，其直方图通常表现出多峰特征。需要指出的是，这种统计规律虽具有普遍性，但在特定成像条件可能出现分布特性偏移（如环境光变化、拍摄时段差异等）。因此在工程实践中，需要结合具体场景参数对直方图特征进行动态解析与适配。

3.3 红外图像数据集介绍

3.3.1 KAIST

韩国先进科学技术研究所(Korea Advanced Institute and Science and Technology, KAIST)红外图像数据集是一个广泛用于计算机视觉和机器学习的重要资源之一，其专门针对红外图像处理，包含了一系列高分辨率的红外图像样本。其部分数据集示意图如图 3-4 所示。

图 3-4 KAIST 数据集示意图

该数据集特点在于其多样性，包含了静态、动态、夜间、白天等场景图像，其多光照条件下的红外图像有助于模型能够适应各种复杂环境变换，通过该数据集，研究人员可以很好地测试算法对红外光谱的理解能力，尤其在低可见度和高温环境中。

其包含校园、街道、农村各种复杂交通环境下的场景，图片尺寸为640×480，分为训练集 50187 张图片和测试集 45141 张图片，标签包含 person、people、cyclist 三个类别，容易区分的个体标为 person，较模糊则为 people，cyclist 则为骑行者。

3.3.2 FLIR

前视红外(Forward Looking Infrared, FLIR)是一种非接触式的热成像技术，在各种复杂条件下，都可以很好地采集图像。FLIR 数据集包括大量实时静态图像，由 Flir Systems 公司采集，于 2018 年发行，其图像数据包括：

热度映射：即每个像素都有其特定的温度值，研究人员可以通过热度映射来分析物体的温度差异，从而识别热源。

高维度信息：其图片为二维数组，每个元素代表一个区域的热辐射强度。

多样化的应用场景：其数据集中包含工业检测、目标追踪、温度预测等应用场景，使得研究人员可以更好地运用在不同复杂场景。

其包含完全漆黑、烟雾、恶劣天气等场景，共 14000 余张红外图像，数据标签包含行人、骑行者、动物、车辆。分为训练集 8862 张图像、测试集 1366 张图像和视频图像 4224 张图像。其部分数据集部分图像如图 3-5 所示。

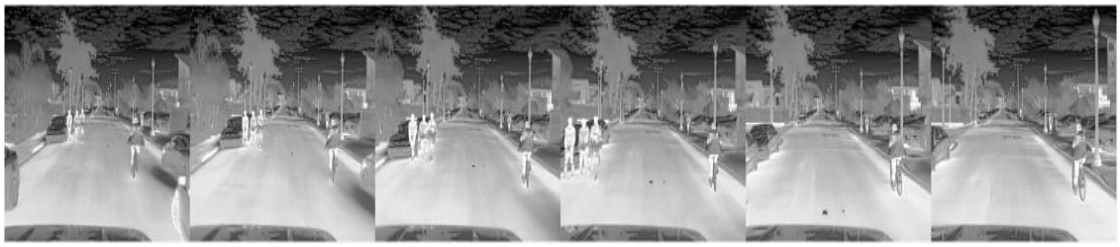


图 3-5 FLIR 数据集部分图像

3.4 夜间道路场景行人检测数据集的制作

红外图像的空间分辨率相较于可见光图像而言较低，其中最主要原因在于红外光的波长较长，需要更长的时间才能积累足够的能量，从而产生一个可靠信号。红外图像细节纹理较模糊，需要捕捉特征纹理细节较难。红外图像的边缘特点较可见光而言不明显，由于红外图像的特殊性质和复杂性导致边缘检测难。

红外图像受限于其成像机制，在视觉表征层面呈现固有低对比度、纹理模糊化及信噪比劣化等特征。这些物理约束直接导致在后续处理流程中，针对目标物的特征提取、目标识别和运动追踪等环节均存在精度衰减问题。为解决此类技术瓶颈，通过对比度重映射算法改善图像动态范围、运用边缘增强技术强化目标轮廓清晰度等预处理操作已成为必要技术路径，充分凸显了红外图像特征增强技术在计算机视觉任务中的基础支撑作用。

3.4.1 红外图像细节增强

通过分析红外成像系统的技术特性可以发现，其核心挑战源于宽动态范围

（如地表-天空超过 300K 的温域跨度）与目标-背景微温差（常低于 1K）之间的矛盾。工程实践中采用 14bit 以上模数转换器件实现原始信号的高精度捕获，可确保百万级灰度层次的物理量精准量化。然而受限于通用显示设备 8bit 色深标准，需通过动态范围压缩技术进行数据适配，此过程易引发两类信息损失：（1）大梯度温度区间的线性特征畸变；（2）微温差目标的对比度衰减。当前主流的压缩算法存在显著局限性：AGC 线性映射虽能保持全局灰度分布，但在处理占空比低于 5%的小目标时会导致有效信号淹没；HE 非线性增强虽提升整体对比度，却会破坏场景辐射能量的物理对应关系。近些年，随着国内外对该问题的研究提出一些算法，其中 DDE 算法脱颖而出，DDE 是一种高级非线性图像处理算法。由于在人体视觉系统和机器扫描仪器中，人类只能识别图像中 128 级灰阶（最亮与最暗之间的亮度变换度），而对于红外热成像仪来说，其探测范围可达 15000 级灰阶以上。FLIR 公司研发的数字细节增强(Digital Detail Enhancement, DDE)算法突破传统框架，通过构建多尺度特征感知模型，在保持背景辐射量连续性的同时，实现特定温阈（如±2K）内小目标的对比度定向提升，可以在图像处理过程中保留高动态场景中图像的细节部分，保证不丢失信息。其处理流程如下。

（1）对于分离低频部分和高频部分：使用特殊的滤波器（引导滤波器）将图像分成低频部分（背景层）和高频部分（细节层）。

（2）对于低频部分进行灰度增强：通过对低频部分使用灰度增强算法，提高低频部分的对比度和视觉感知。

（3）对于高频层进行细节增强和噪声抑制：高频层中含有图像的细节信息，可以利用边缘增强等算法来增强细节抑制噪声。

（4）动态范围调整：根据图像整体的动态范围，对低频部分和高频部分进行动态范围内的调整等操作，从而达到将原本动态范围较高的图像信息映射到 8bits 的范围内。

（5）合成输出图像：将增强后的低频部分和高频部分重新合成 8bits 范围的输出图像，从而显示大动态温差和目标局部细节信息。

3.4.2 评价指标

通过对使用 DDE 算法之后的图像与原始图像进行对比实验，本研究建立跨模态质量评估框架验证 DDE 算法效能，构建三阶量化评估体系。包括均方误差(Mean Square Error, MSE)、峰值信噪比(Peak signal-to-noise ratio, PSNR)、结构相似性(Structural Similarity, SSIM)。

MSE 是衡量图像质量的重要指标之一，MSE 表示原始图像与处理后图像之间的差异，MSE 越小，说明对图像质量越好，其原理为真实值与预测值之间的差值平方的平均值，其公式如式(3-2)所示：

$$MSE = \frac{1}{M \cdot N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [I(i,j) - K(i,j)]^2 \tag{3-2}$$

其中，M 为图像 I 的像素总数，N 为图像 K 的像素总数。

PSNR 常用于表示信号最大可能功率与影响其表示保真度的破坏噪声功率之间比值，在图像处理上通常用来量化受损图像与重建后图像质量之间的误差，PSNR 越大，说明图像增强效果越好，其计算公式如式(3-3)所示：

$$PSNR = 10 \log_{10} [\frac{(2^n - 1)^2}{MSE}] \tag{3-3}$$

其中，n 为图像灰度比特数。

SSIM 常用于评估两幅图像相似度的指标，用于衡量两幅图像的相似性，同时也被用于衡量模型生成图像的真实性，在图像处理中表示原始图像和处理后图像之间结构上的相似性，SSIM 值越接近于 1，说明图像增强效果越好，其计算公式如式(3-4)所示：

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{3-4}$$

其中，μ 、σ 分别为红外图像块之间的均值、方差，σ_{xy} 为红外图像块 x 和红外图像块 y 之间的协方差，C_1、C_2 为常数，通常为 0。

3.4.3 对比实验

采用不同方法对数据集中部分夜间场景红外图像进行增强处理，获取 MSE、PSNR、SSIM 评价指标，结果如表 3-1 所示：

表 3-1 图像处理效果对比实验

Evaluation indicators	DDE(ours)	文献[57]	文献[58]
MSE/%	0.705	1.126	0.989
PSNR/dB	51.46	48.73	45.05
SSIM	0.967	0.882	0.898

由表可知，本文采用的算法进行红外图像增强 MSE 较其他对比算法较低，为 0.705%，PSNR 较高，为 51.46dB，SSIM 效果较好，为 0.967。而文献[57]和文献[58]的方法 MSE 虽然较高，分别为 1.126%、0.989%，且其 PSNR、SSIM 达到了红外图像处理的要求，但均低于本文图像增强算法。由此表明本文算法红外图像增强效果较好，且处理后红外图像分辨率较高。最终红外图像增强效果对比图如图 3-6 所示。

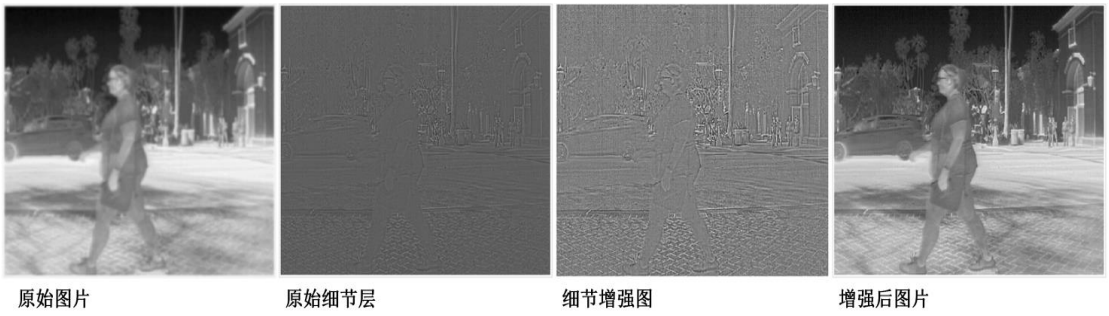


图 3-6 红外图像增强效果对比图

3.4.4 数据集标注

本文基于夜间道路行人检测数据集研究中使用的数据来自 FLIR 公司提供的热成像数据集，该数据集包含 14000 张图像，其中 10000 张来自短视频片段，另外 4000 张 BONUS 图像来自一段 140 秒视频图像捕获刷新率。

本文使用的是 YOLO 系列算法，在训练前需要对数据集进行标注处理，标注行人的真实框，采用 LabelImg 标注工具，形成的标注文件格式是 YOLO 形式，标注时只需标注行人，效果图如图 3-7 所示：

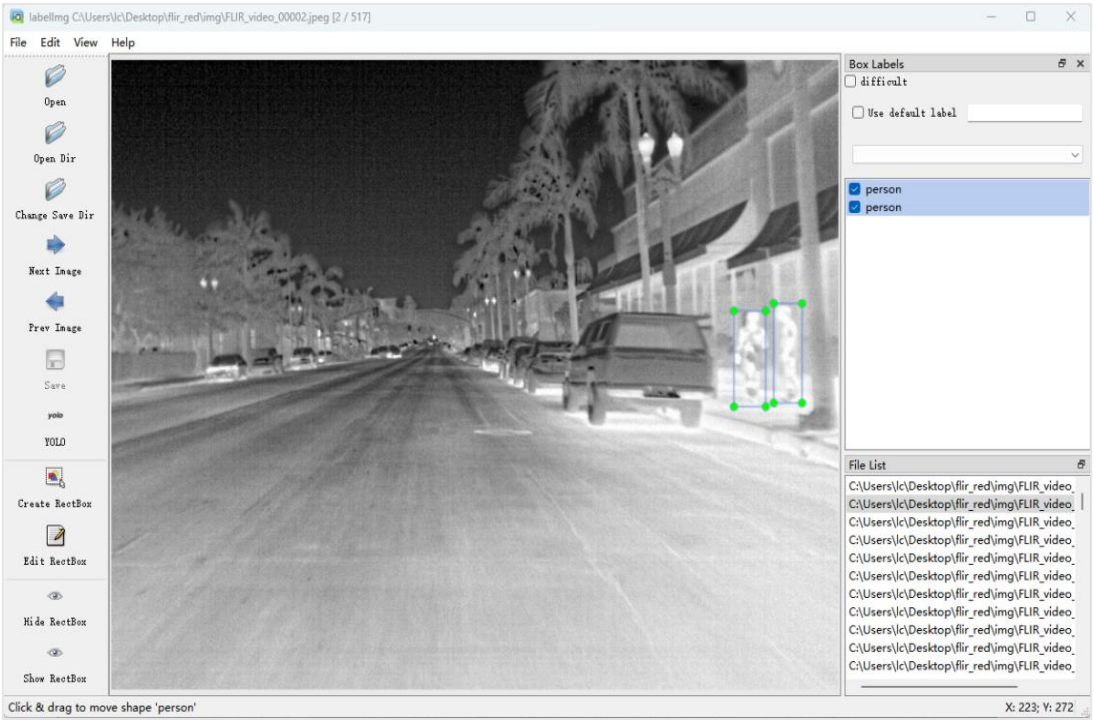


图 3-7 LabelImg 标注框

在标注数据集时，在终端或者命令行中输入 labelimg 命令，即可打开标注软件，选择数据集目录在 labelimg 中，点击 Create RectBox 随后利用鼠标对行人进行矩形框标注，为了区别红外图像行人数据集和红外-可见光行人数据集，由于在可见光严重不足的情况喜爱骑行者较少，所以统一标注默认为行人，标注完毕后标注命名框为 person，在红外-可见光图像行人数据集中，由于夜间存在部分可见光的情况下，道路行人包括步行的行人与骑行的行人，所以分别标注步行者和骑行者，为了区别单纯的红外图像数据集中行人数据，将其步行者命名为 pedestrian，标注完毕后分别标注命名框为 pedestrian 和 bicycle。

3.4.5 划分夜间行人检测数据集

根据本文实验架构分别划分出夜间场景下红外图像行人检测数据集与夜间场景红外-可见光图像行人检测数据集。数据集建立完毕后划分为训练集和验证集。

（1）红外图像行人数据集的划分

由于 FILR 数据集包含白天与夜间可见光与红外图像，所以需要从中筛选出符合夜间条件下的红外图像作为实验红外图像行人检测的数据集。最终通过从中

筛选出 2063 张夜间红外图像作为红外图像行人检测实验数据集，为了确保实验的鲁棒性，增加 200 张红外无人图像作为实验数据集中负样本，以此提升算法的鲁棒性。

（2）红外-可见光图像行人数据集的划分

由于红外图像行人检测数据集中没有专门针对夜间具有可见光的场景，所以实验需要重新划分数据集，其包含夜间可见光图像及其对应场景的红外图像。通过筛选同一夜间道路场景下 7381 张红外图像和 7381 张可见光图像作为红外-可见光行人检测实验数据集。

同样为了确保实验的鲁棒性，增加 750 张同一场景红外-可见光无人图像作为实验数据集负样本，以此提升算法的鲁棒性。

本文实验数据集部分正负样本图如图 3-8 所示。

（a）部分正样本

（b）部分负样本

图 3-8 数据集部分正负样本

3.5 本章小结

本章首先介绍了红外成像技术理论包括红外成像原理及红外图像特点，然后介绍部分开源红外图像数据集。针对红外图像的特性，对红外图像进行特征细节

增强，以此来提高目标与背景的对比度，同时提高红外图像目标的清晰度，通过使用常见的图像性能评价指标对处理后的图像与原始图像进行对比实验，从而验证算法的有效性。然后在此基础上对行人目标进行人工标注，建立了后续实验所需要的两个数据集，分别为红外图像行人数据集和红外-可见光图像行人数据集，最后划分为训练集和验证集。

第4章 基于红外图像的行人检测算法研究

4.1 引言

行人检测作为自动驾驶和辅助驾驶发展过程中重要的一点，由于车载设备硬件条件有限，所以兼顾检测的实时性和准确性是至关重要的。同时，由于可见光的影响导致行人检测方法各不相同。传统行人检测一般在可见光较足的情况下采用可见光探测器采集图像并进行检测，然而当可见光不足的情况下，由于光照条件的缺失，通常无法达到理想的检测效果。当然除了光照条件的影响，目标的距离远近也会带来影响，距离较为远的目标通常由于容错率，导致目标纹理细节部分缺失，信噪比低的情况，从而影响到检测的精度，所以这是当前行人检测所面临的重大挑战。在可见光不足的情况下通常采用红外成像仪采集图像，红外成像仪根据待测目标的红外电磁波，待测物体与周边温差的原理生成图像。通常在温度监测、军事制导等领域大显身手。传统的夜晚红外图像目标检测算法通常根据人为设计的特性。

为解决目标检测实时性与准确性不能同时兼顾等问题。本文通过改进相关行人检测算法，提出一种基于 YOLOv8 改进的红外图像行人检测算法 RT_YOLOv8。进行主干网络替换实验并筛选出 RepGhostNet，通过结构重新参数化技术实现特征的隐式重用实现模型体积的降低；引入 Triplet Attention，消除通道和权重之间的间接对应；使用可变形卷积重新构建 C2f(CSP Bottleneck with 2 convolutions)模块，提高模型对行人目标大小不一的识别能力，提高检测精度；使用 PConv 替换 Detect 模块中的部分原有 Conv，进一步实现轻量化。

4.2 YOLOv8 算法简介

在目标检测领域，YOLO 一直是一种突破性算法。从 YOLOv1 到 YOLOv8 各种版本，其中最受欢迎的当数 YOLOv5 和 YOLOv8，在各自领域独领风骚。

YOLOv8 是一个 SOTA 模型，和 YOLOv5 一样具有 n、s、m、l、x 模型，YOLOv8 网络结构如图 4-1 所示，分为 Backbone、Neck、Head 三部分，Backbone

用于特征提取，由 Conv、C2f、SPPF(Spatial Pyramid Pooling)模块组成，Conv 对输入数据进行卷积、BN(Batch Normalization)、SiLU(Sigmoid Gated Linear Unit)操作；C2f 根据输入数据不同尺寸调整通道数，降低模型参数，提升收敛速度；SPPF 对输入大小不同的特征图转换为固定大小的张量。Neck 端负责进行多尺寸特征融合，结构为 PANet，PANet 包含 FPN 和 PAN 两部分，两部分的结合实现上下特征信息之间的传递，再通过上采样、通道拼接融合输出到 Head 端。Head 端采用目前流行的解耦头结构，将回归分支和预测分支分离，同时也从 Anchor-Based 换成了 Anchor-Free，加速了模型的收敛速度。

YOLOv8 的主题思想是将目标检测问题划归为回归问题。通过将输入图像分成 S×S 的网格，对于每个网格算法会预测其中是否存在目标的位置，并且通过判断输出一组边界框，每个边界框分别由中心点的坐标和边界框的宽度和高度来确定，并且每个边界框会分配一个置信度分数，从而判断该边界框中是否包含所需要检测的目标。

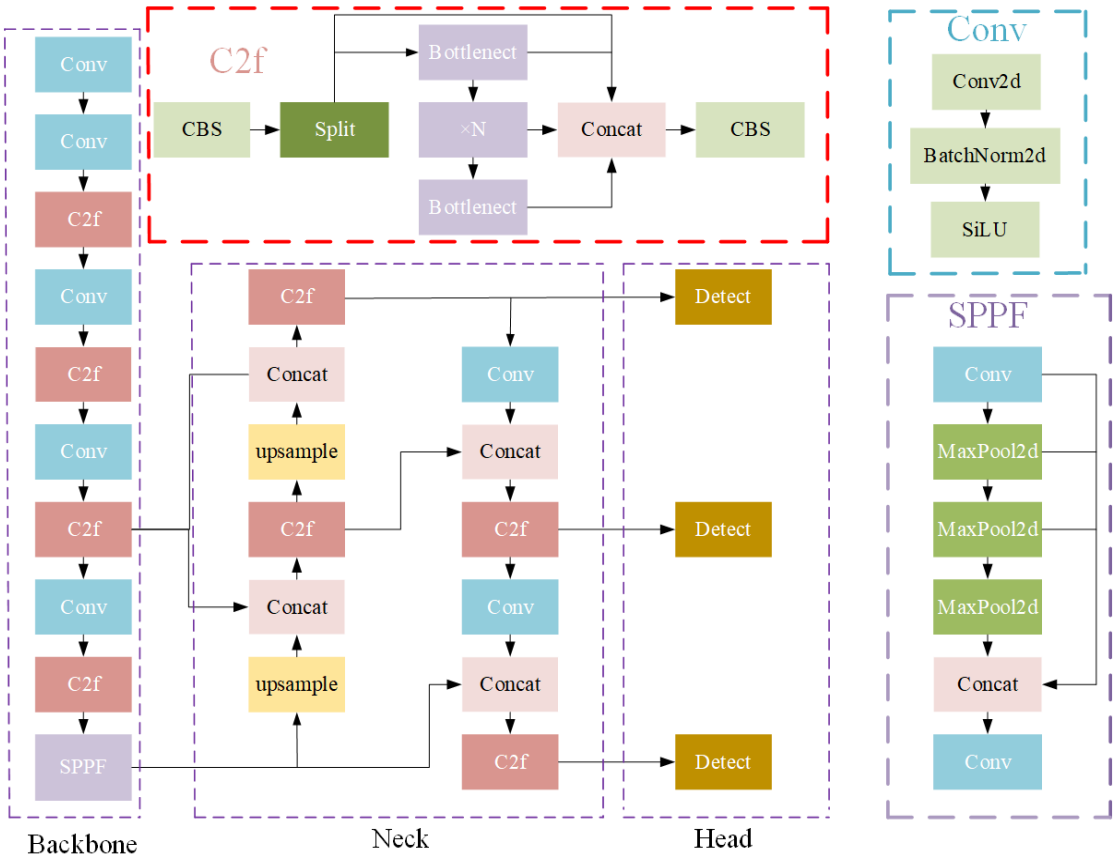


图 4-1 YOLOv8 网络结构

4.3 红外图像行人检测网络结构设计

本文从红外图像行人检测中实时性与准确性平衡的角度出发对 YOLOv8 基准模型网络进行改进，改进后模型网络结构图如图 4-2 所示。Backbone 端对数据进行特征提取；Neck 对特征提取后数据进行特征融合；Head 端负责数据的输出。改进算法集中关注 YOLOv8 模型中 Backbone、Neck 端中的 C2f 模块、Head 端中的 Detect 模块。为解决当前行人检测技术面对红外图像检测准确性存在误差较大、硬件设备要求较高等问题，本文针对相关问题，提出一种改进 YOLOv8 的检测网络模型 RT-YOLOv8。使用轻量级 CNN 模块 RepGhostNet1.0 替换 YOLOv8 主干网络，只保留空间池化金字塔特征(SPPF)来控制一个自适应尺寸的输出。该模块通过结构重新参数化技术实现特征的隐式重用；引入 Triplet Attention，在新引入的主干网络 SPPF 模块后添加，该注意力机制计算量几乎微乎其微，并且通过通道、高度和宽度三个维度的交互从而更好地捕捉特征信息；基于可变形卷积构建了 C2f_DVNv2 模块来替换 Neck 端中的 C2f 模块，以提高模型对不同形状物体识别能力；提出一种新的 PConv(partial convolution)替换 Head 端中的 Detect 模块中部分原有 Conv，通过减少冗繁内存访问从而达到降低计算量的目。

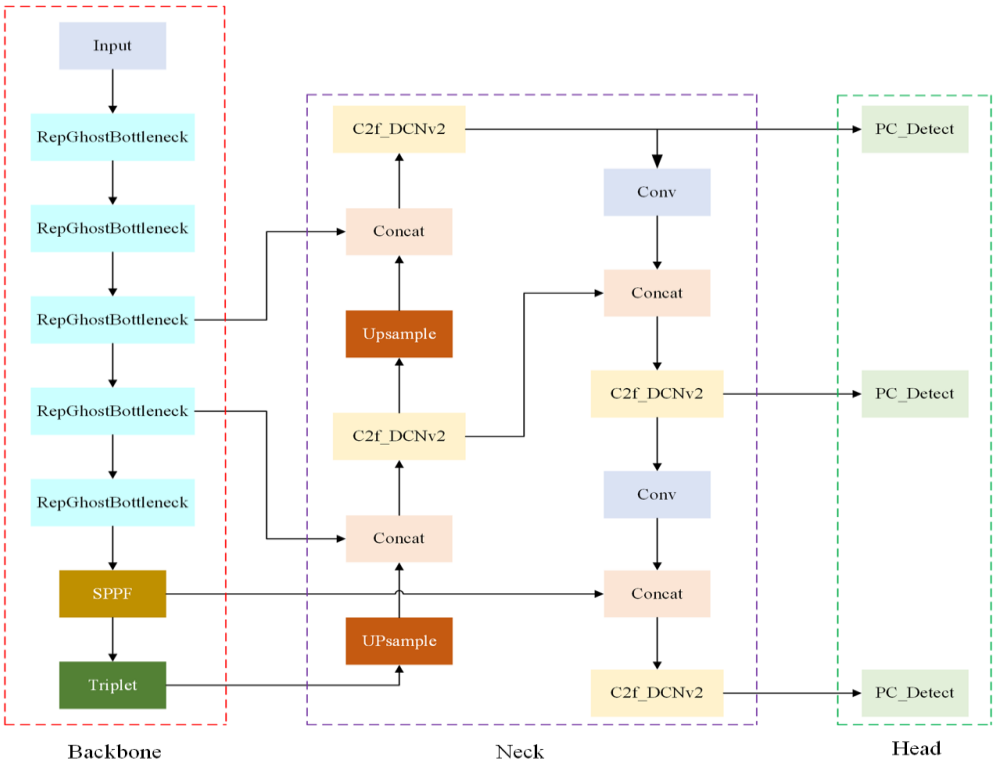


图 4-2 RT-YOLOv8 网络结构

4.3.1 RepGhostNet 替换

YOLOv8 主干特征提取网络 FasterNet 带来较大的参数量，导致识别速度较慢。因此本文采用 RepGhostNet[57]网络结构替换，以实现网络的轻量化，提高检测的实时性。RepGhostNet 是一种通过结构参数化技术开发的硬件高效的 RepGhost 模块，从而达到特征的隐式复用。特征复用一直是当前轻量级卷积神经网络中必不可少的一种技术。在卷积操作中 Concat 操作不需要计算量的，然而在硬件设备中是需要考虑计算成本的。RepGhost 模块则通过结构重参化技术达到特征重用，并不使用 Concat 操作，通过替换 Ghost 模块中低效的串联算子，实现隐式复用。RepGhost 模块中将原有 Ghost 模块中 Concat 操作替换为 Add 操作，并在使用 Add 操作后的激活函数来满足重新参数化结构的原则。在 RepGhost 模块的基础上设计 RepGhost Bottleneck 和 RepGhostNet。

RepGhostNet 的网络结构如图 4-3 所示。通过使用多个 1×1 卷积层和平均池化层(AvgPool)对输入通道进行处理，最终输出的预测得以实现。根据输入数据尺寸将 RepGhost Bottleneck 分成五组不同尺寸数据，并且每组中最后一个 Bottleneck 设置 stride 为 2。

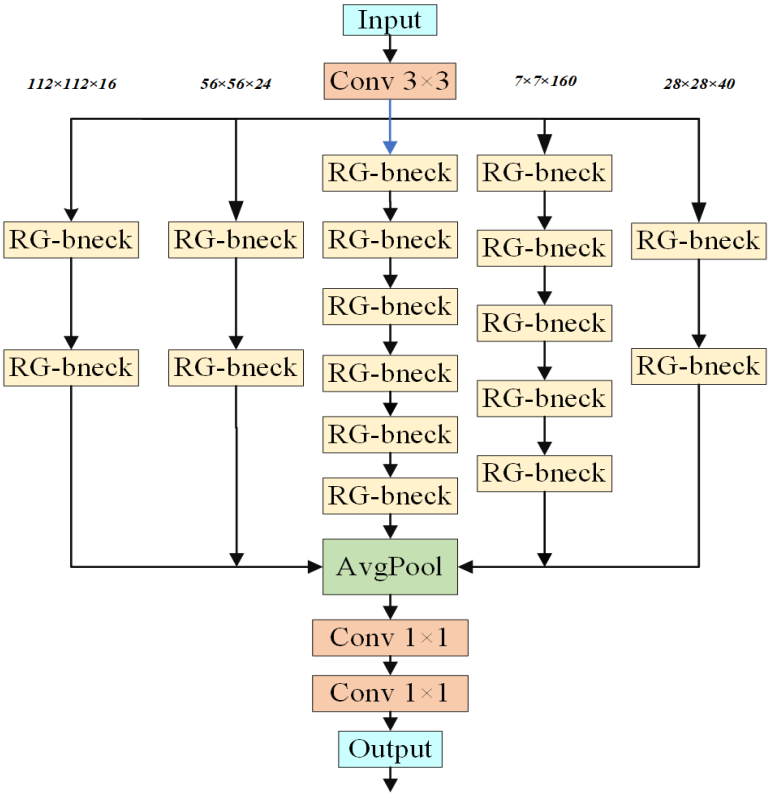


图 4-3 RepGhostNet 网络结构

RepGhost 模块重参化过程如图 4-4 所示，首先在归一化(BN)模块后应用可变形卷积进行特征融合，随后将该融合结果与分支路径上的 BN 模块输出进行二次整合后输出。RepGhost 模块通过对输入图片进行卷积操作，拓展特征维数后将分支生成的特征图相加再通过激活函数 ReLU 从而实现重参化。RepGhost 模块在进行推理时仅有规则形态的卷积层与激活函数 ReLU，因此具有较高的效率。

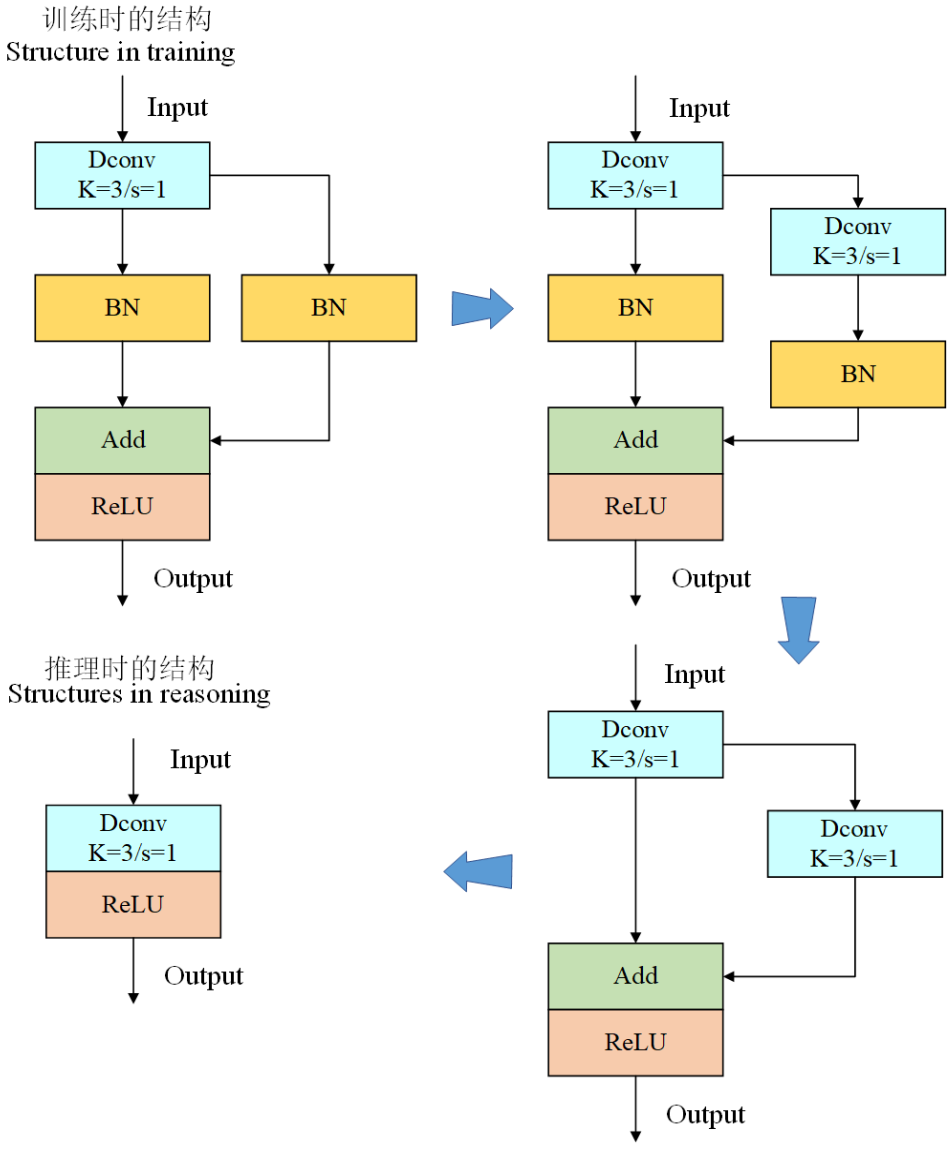


图 4-4 RepGhost 模块重参化过程图示

4.3.2 Triplet 嵌入式设计

为了增强 YOLOv8 模型的表征能力，增强有效信息关注度，在 Backbone 末端 SPPF 模块后增加 Triplet[60]注意力机制。该注意力机制是一种几乎无参的注意

力。对于输入张量，Triplet 注意力机制通过旋转后经过残差变换从而建立各种维度之间的依存关系。其优点在于计算量几乎没有，并且通过通道、高度和宽度三个维度的交互能够更好地捕捉到更多特征信息。

Triplet attention 网络结构如图 4-5 所示，分支 1：输入特征经过通道池化，再经过一组 7×7 卷积，最后通过激活函数生成空间注意力权重。分支 2：通道 C 和宽度 W 维度交互捕获分支，输入特征经过转置函数，维度信息转为 H×C×W，之后在 H 维度上经过 Z-Pool，再经过 Batch Norm。最后经过转置函数形成 C×H×W 的维度。分支 3：通道 C 和高度 H 维度交互捕获分支，输入特征经过转置函数，转成 W×H×C 维度，然后在 W 维度上经过 Z-Pool，再经过 BN 模块，最后由转置函数变成 C×H×W 维度。最后对以上三个分支输出特征进行相加求平均。其中 Z-Pool 为对输入进行 *MaxPool* 与 *AvgPool* 操作，最后输出 2×H×W。Z-Pool 公式如式(4-1)所示：

$$Z-Pool=[MaxPool,AvgPool] \tag{4-1}$$

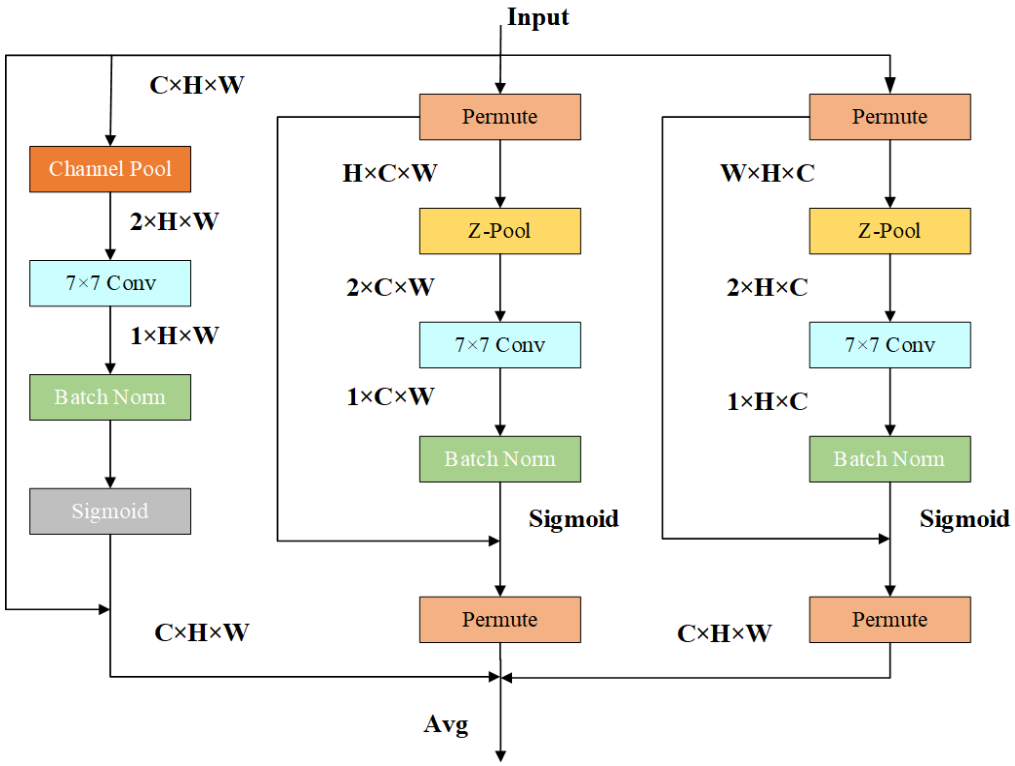


图 4-5 Triplet attention 网络结构

4.3.3 重新构建 C2f 模块

传统卷积操作具有固定的感受野和权值分布，对于形状不同的复杂目标会受

限。然而在C2f模块中引入可变形卷积DCNv2[61-63]后则能很好地改变检测效果。传统卷积无法适应不规则尺寸图形，而可变形卷积通过引入一个偏移量达到任意采样的目的。可变形卷积过程如图 4-6 所示。首先常规卷积根据输入特征生成offset，根据生成后的offset生成像素点，每个像素点转换为随机分布的9个像素点，新生成的像素点对输入特征进行采样后得到特征图。特征图再与卷积核相乘求和得到所需要的卷积结果。

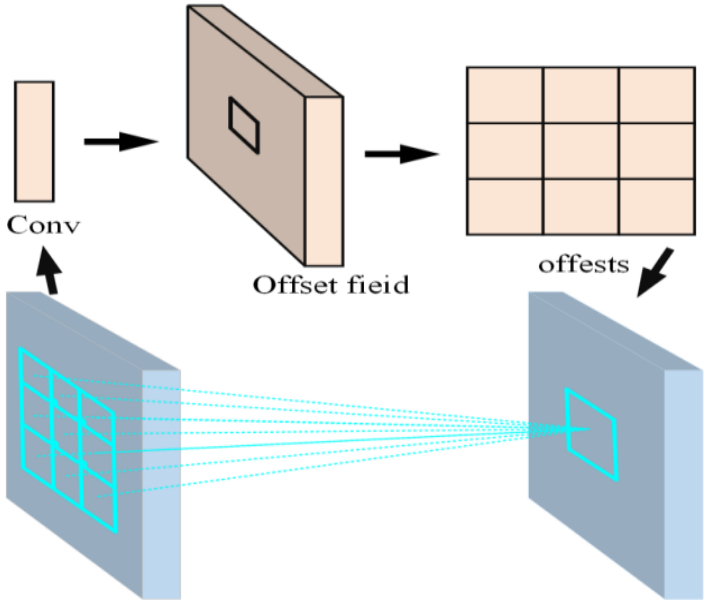


图 4-6 可变形卷积过程图

标准卷积的特征矩阵公式如式(4-2)所示：

$$y(p)=\sum_{i=1}^{k} w_k \cdot x(p+p_k) \tag{4-2}$$

其中 w 为卷积核， p_k 为卷积核的位置，k 为卷积核采样位置个数。

在式(4-2)中加入可学习偏移量可获得 DCNv1 特征矩阵如式(4-3)所示：

$$y(p)=\sum_{i=1}^{k} w_k \cdot x(p+p_k+\Delta p_k) \tag{4-3}$$

其中， Δp_k 为新增二维偏移矩阵。

对式(4-3)每个采样点加入权值项，消除无关区域，获得 DCNv2 特征矩阵如式(4-4)所示：

$$y(p)=\sum_{i=1}^{k} w_k \cdot x(p+p_k+\Delta p_k)\cdot \Delta m_k \tag{4-4}$$

其中， Δm_k 为归一化后的调节项。

在神经网络中，待检测目标形状各异，所以本文只对 YOLOv8 网络结构中

的 Neck 端的 C2f 模块进行改进，将 C2f 模块中 Bottleneck 结构中的部分 Conv 替换为可变形卷积，改进后 C2f_DCNv2 结构如图 4-7 所示。以张量(h, w, c)作为输入，通过 1×1 的卷积后由 Split 拆分为两份(h, w, 0.5c)，一份直接输入到 Concat，另一份输入到 Bottleneck_DCNv2 序列中，经过 n 个 Bottleneck_DCNv2 后输出到 Concat，共计 n+2 个张量(h, w, 0.5c)。0.5(n+2)c_out 再通过 1×1 卷积得到 c_out。由于 C2f 模块位于骨干网络中，所以 Bottleneck_DCNv2 具有残差边，在内部分别进行 shortcut 操作和经过 Conv、DCNv2 后输入到 add。

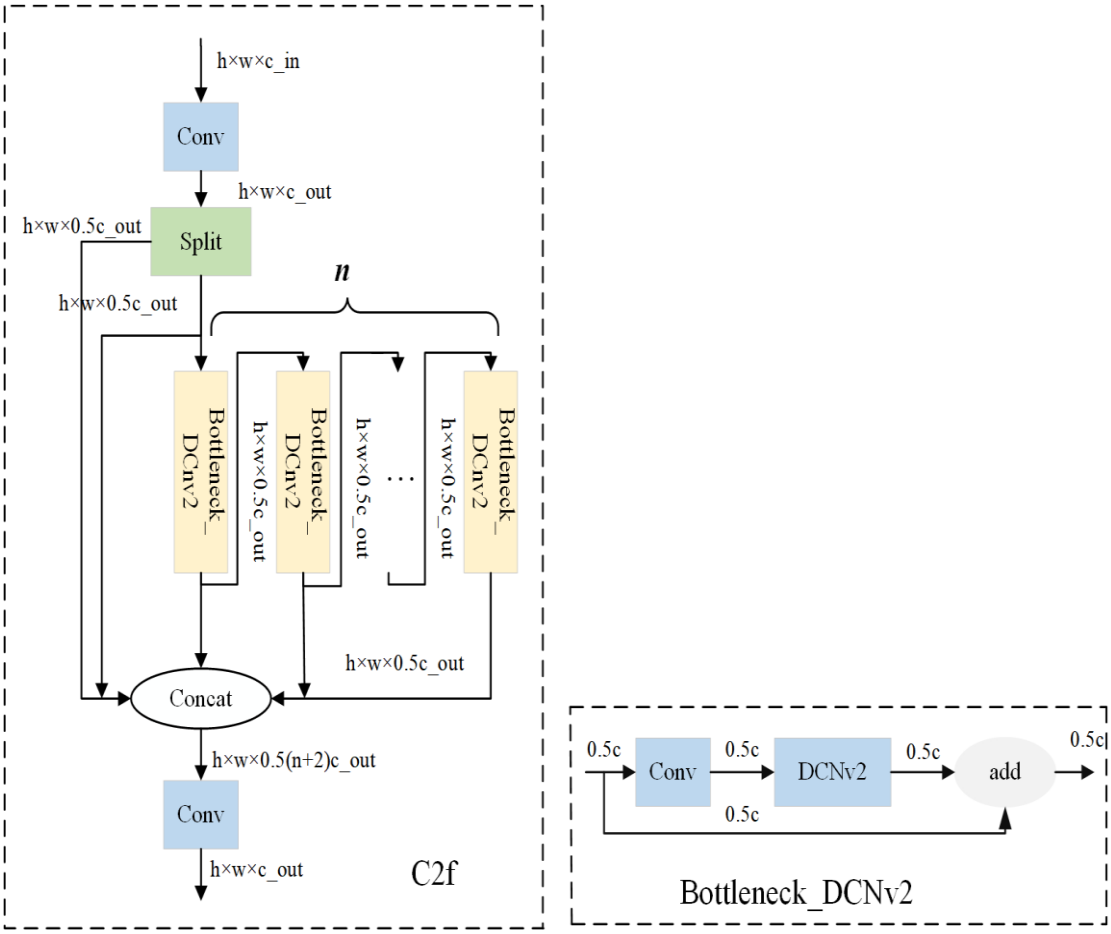


图 4-7 C2f_DCNv2 结构

4.3.4 Detect 模块的改进

通过对神经网络的研究，多数快速神经网络集中对浮点运算(FLOPS)的数量进行降低。然而大部分浮点运算的减少并不会带来很明显的延迟减小，由于常规卷积操作会加剧对内存的访问，为了减少运算内存访问量，在 Detect 模块中提出一种新的 PConv(partial convolution)[64]替换 Detect 模块中的部分原有卷积层。从

而达到减少多余计算量和内存访问量。

图 4-8 中（a）为常规卷积与纵深/分组卷积的结构图。图 4-8 中（b）为 PConv 结构图，将高维数据的特征向量映射到低维数据的过程中存在部分冗余，在部分输入通道中使用 Conv，其相比于常规卷积具有更低的计算量。图 4-8 右侧为 Detect 结构图，在 Detect 模块中输入数据先经过一个 3×3 的卷积再通过 PConv 操作后经过一个 1×1 的卷积，最后分别结算 Bbox.Loss 和 Cls.Loss。

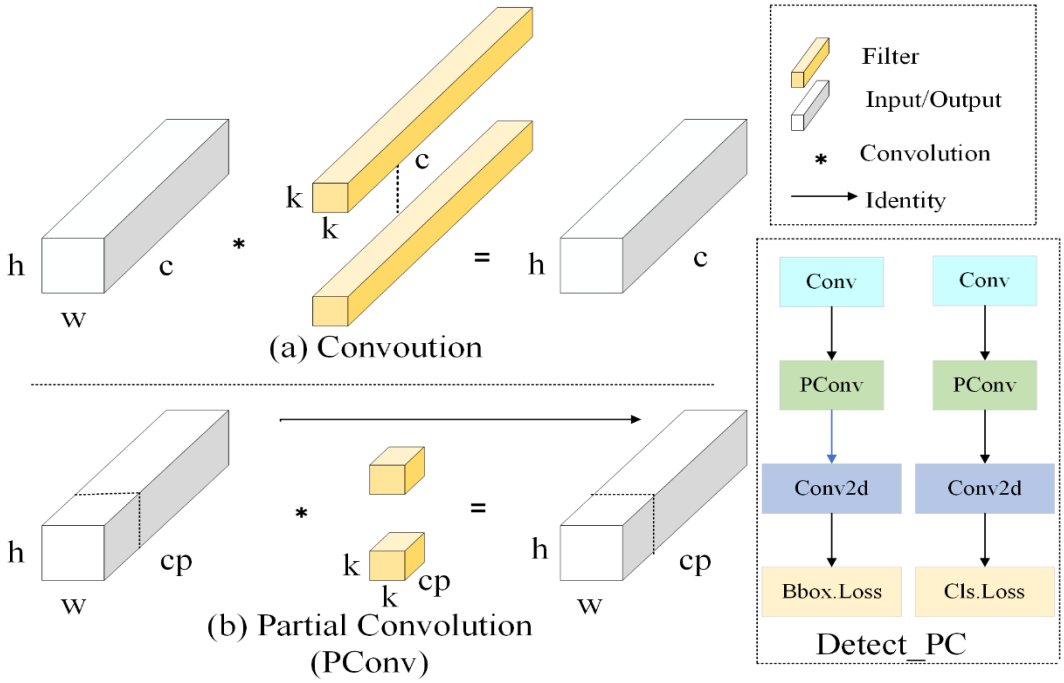


图 4-8 PConv 及 Detect_PC 结构

4.4 实验结果与分析

为了验证本文提出的改进方法实际效果，达到分别对 YOLOv8 主干网络、注意力机制和 C2f 模块进行改进，并优化调整参数，通过实验对比各种改进方法的效果。实验结果最优数据加粗标注，次优数据用下划线标注。

4.4.1 实验配置与参数

为了验证本章所提出红外图像行人检测实时性与准确性的研究算法的有效性，本章采用第三章所建立的红外图像行人数据集上进行实验，改进网络在红外图像行人数据集上训练的损失函数变化如图 4-9 所示，实验结果显示当迭代次数

达到 100 轮次时，损失函数曲线进入平稳状态，表明模型已达到优化稳定阶段，本章最终选定 100 轮次为模型训练周期。

实验在系统 ubuntu22.04 系统环境下进行，硬件设备为显卡 8G 显存的 NVIDIA GeForce RTX1070，软件环境为 python3.8 使用 torch1.13.1 框架，设置超参数 learning rate 为 0.01，batch size 为 16。

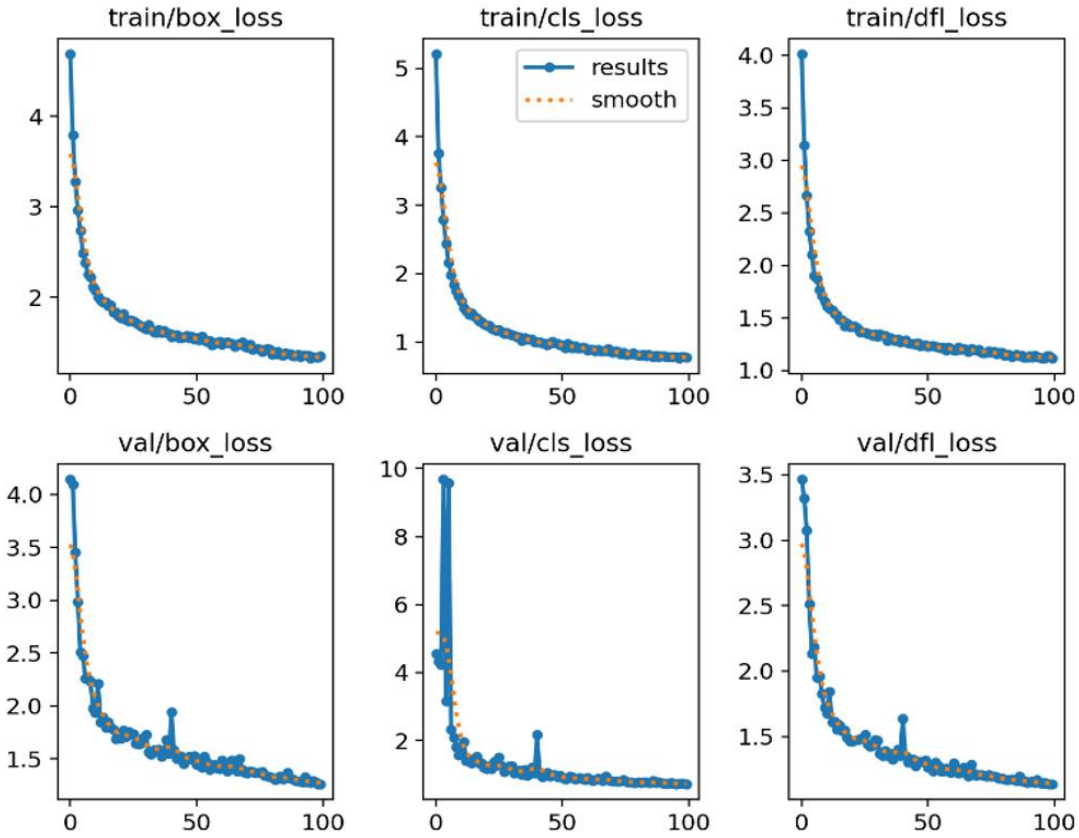


图 4-9 损失函数变化图

4.4.2 主干网络改进实验

将本文所选主干网络 RepGhostNet 与常用主干网络进行对比实验。分别在本文建立的红外图像数据集上进行实验进行模型改进性能对比。改进主干网络后的模型性能对比如表 4-1 所示，使用 RepGhostNet 模型参数量最低，相比于使用 CSPDarknet 的 YOLOv8 模型，原模型和改进后 mAP@0.5 分别为 91.50%、91.70%，mAP@0.5 提升 0.20%，原模型和改进后参数量分别为 11.47M 和 8.92M，参数量降低 22.23%。说明使用 RepGhostNet 可以很好地降低模型的大小。

表 4-1 改进主干网络后的模型性能对比

Backbone	mAP@0.5 (%)	Parameter (M)
CSPDarknet	<u>91.5</u>	<u>11.47</u>
EfficientViT	90.1	15.29
RevCol	89.4	12.88
RepViT	90.7	25.55
MobileNetV3	87.1	11.87
GhostNetV2	91.1	11.71
RepGhostNet(Ours)	**91.7**	**8.92**

4.4.3 C2f 模块改进实验

分别使用 FasterNet 中的 FasterBlock 替换 C2f 中的 Bottleneck、使用一种新的 building block：DDB(Diverse Branch Block)替换 C2f 中的 Bottleneck 中的 Conv、使用一种轻量级 Vision Transformer 架构 CloFormer 中的具有全局和局部特征的注意力机制添加到 C2f 中的 Bottleneck、使用 CGnet(Context Grided Network)中的 Light-weight Context Guided 改进 C2f、引入可变形卷积 DCNv2 替换 C2f 中部分 Conv。分别在本文建立的红外图像数据集上进行实验进行模型改进性能对比，C2f 模块改进后的模型性能对比如表 4-2 所示。

表 4-2 C2f 模块改进后的模型性能对比

Model	mAP@0.5 (%)	Parameter (M)
C2f	<u>91.5</u>	11.47
C2f+Faster	89.8	<u>8.78</u>
C2f+DBB	91.1	11.47
C2f+CloAtt	91.0	11.93
C2f+ContextGuided	90.9	11.59
C2f+DCNv2(Ours)	**91.7**	**8.02**

由表 4-2 可得，使用 DCNv2 后模型平均精度较之其他模型最高，原模型和改进后参数量分别为 11.47M 和 8.02M，参数量降低 30.08%，在有效降低模型体积的同时还略微提高检测准确率。

4.4.4 消融实验

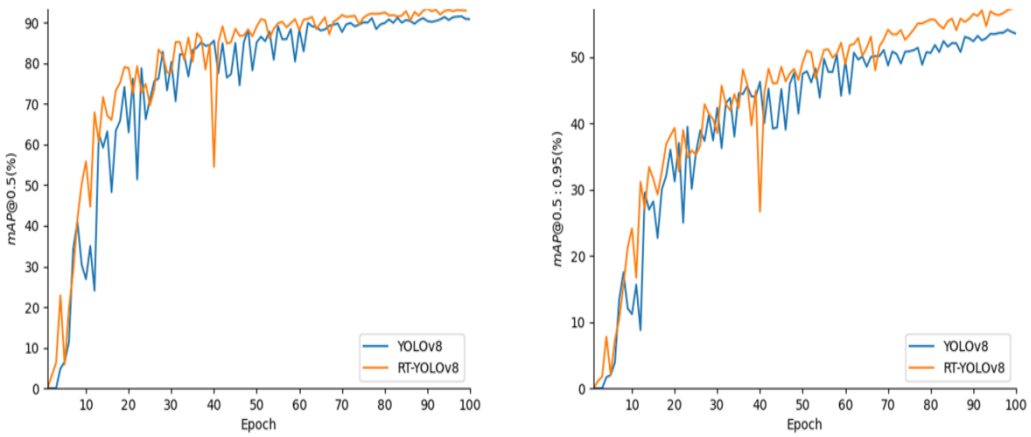
为进一步验证各改进模块之间的组合性能设计消融实验，消融实验在本文建立的红外图像行人数据集上进行。消融实验结果如表 4-3 所示。加粗数据表示当前指标最优值，底部加横线数据表示当前指标次优值。

表 4-3 消融实验结果

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	Parameter (M)	GFLOPs (G)
YOLOv8	**93.4**	83.1	91.5	11.47	8.1
YOLOv8+RepGhost	91.7	84.7	91.7	8.92	6.8
YOLOv8+RepGhost+Triplet	91.7	84.3	91.0	9.02	6.8
YOLOv8+RepGhost+Triplet +Detect_PC	91.3	82.1	90.7	**6.79**	4.2
YOLOv8+RepGhost+Triplet +C2f_DCNv2	92.4	83.2	91.4	9.29	6.6
YOLOv8+RepGhost+Triplet +Detect_PC+C2f_DCNv2(Ours)	**93.4**	**85.0**	**92.9**	7.06	**4.0**

经过消融实验可得出以下结论：主干替换成功的缩减模型参数量和模型计算量。模型在引入 Triplet 注意力机制后，召回率有一定提升。重新构建 C2f 模块和使用 PConv 替换 Detect 模块中部分卷积层后在保证模型准确率没有明显下降的同时，模型参数量和模型计算量有较为明显的下降，通过消融实验表现出模块优化之间的关联性，其组合效果在实验中表现最优。

RT_YOLOv8 模型平均精度提升 1.40%，参数量减少 38.45%，浮点计算量降低 50.62%。如图 4-10 所示模型在第 50 轮时，改进算法平均精度较基准算法有较为明显的提升。验证了 RT_YOLOv8 算法的先进性。



（a）mAP@0.5对比　　（b）mAP@0.5:0.95对比

图 4-10 模型改进前后的 mAP 曲线对比

4.4.5 对比实验

为了更好地确定本文算法的整体性能设计对比实验，分别与当下主流红外行人检测算法进行对比，对比实验在本文建立的红外数据集上进行。模型检测性能对比结果如表 4-4 所示。加粗数据表示当前指标最优值，底部加横线数据表示当

前指标次优值。

表 4-4 主流红外行人检测算法性能对比

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	GFLOPs (G)
YOLOv5n	91.4	84.4	91.0	15.8
YOLOv8n	93.4	83.1	91.5	8.1
YOLOv8s	91.8	84.8	92.0	28.4
YOLOv8m	90.9	84.2	92.2	78.7
YOLOv8l	92.4	84.4	92.8	164.8
YOLOv8x	91.4	88.2	**93.1**	257.4
RT-DETR	74.2	70.8	78.1	103.4
Faster-RCNN	34.7	59.6	59.3	369.7
SSD	92.3	42.5	68.8	60.7
RetinaNet	**97.3**	61.0	74.0	145.3
MobileNetV3-YOLOv8	87.3	78.5	87.1	6.1
CPCA-YOLOv8	91.8	84.4	91.8	8.3
MPCA-YOLOv8	90.2	**93.7**	91.4	8.1
SE-YOLOv8	93.3	83.8	91.8	8.1
BiFPN-YOLOv8	91.1	83.8	91.3	7.1
RT_YOLOv8(Ours)	93.4	85.0	92.9	**4.0**

在数据集上实验可知：从 mAP@0.5 上看，实验对比算法 mAP@0.5 最高为 YOLOv8x 的 93.10%，改进算法的 mAP@0.5 为 92.90%，在对比红外行人检测算法中数据表现次优。从模型计算量和模型参数上看，对比算法次优为 MobileNetV3-YOLOv8 的计算量为 6.1G，改进算法模型浮点计算量为 4.0G，在对比红外行人检测算法中数据表现最优。在对比实验中可以看出，虽然部分算法精度较高，然而模型计算量也随之上升，不利于搭载在算力较低的车载设备上。本文 RT-YOLOv8 算法在保证算法体积较小的同时检测精度也不逊色于其他算法，达到了兼顾实时性与准确性的目标。

为了进一步证明改进算法的可靠性，在公开红外行人检测数据集 KAIST 上进行原 YOLOv8 和 RT_YOLOv8 对比实验。实验结果如表 4-5 所示。

表 4-5 KAIST 数据集测试结果

Model	Precision (%)	Recall (%)	mAP@0.5 (%)	GFLOPs (G)
Original YOLOv8 algorithm	89.4	84.0	91.5	8.1
RT_YOLOv8(Ours)	92.3	83.1	92.0	4.0

在公开红外行人检测数据集上实验得出：随着训练样本数的增加，模型还能在保证较高精度的同时兼顾较低的模型体积，相比于原算法，本文改进后的算法达到了预期的效果，满足兼顾实时性与准确性的目标。

4.4.6 实验结果可视化分析

最终检测结果效果如下，未经过检测的原始图像如图 4-11 所示，YOLOv8n 算法检测结果图如图 4-12 所示，RT-YOLOv8 算法检测结果图如图 4-13 所示。

图 4-11 原始图像

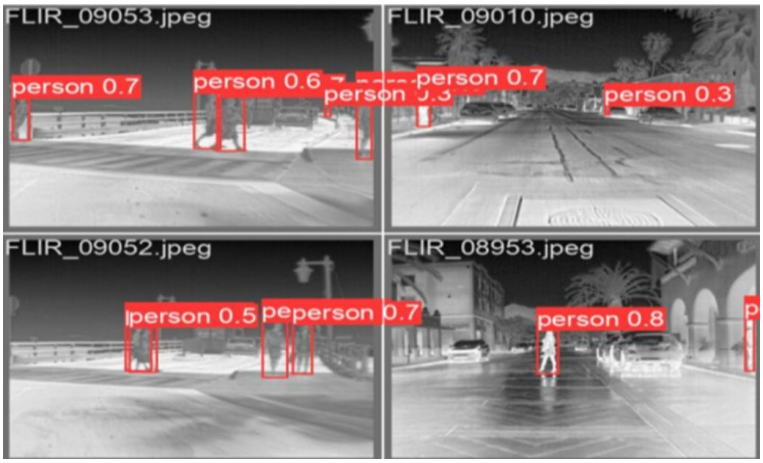


图 4-12 YOLOv8 检测效果图

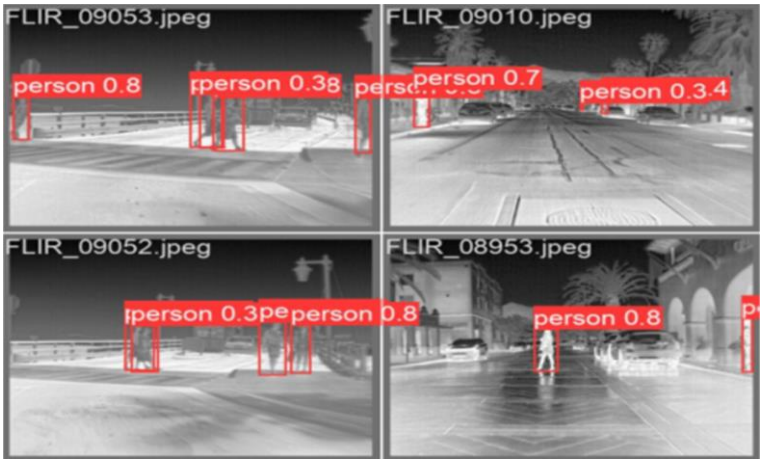


图 4-13 RT_YOLOv8 检测效果图

从图 4-12、图 4-13 对比中可以看出，YOLOv8 在面对远处较小的检测对象时，存在漏检、准确率较低等情况。结果证明 RT-YOLOv8 在准确率和漏检率上都优于原算法。因此可得出结论：RT-YOLOv8 算法更具有实用性，对于红外图像行人检测能够很好地兼顾准确性与实时性。

4.5 本章小结

本文提出一种改进 YOLOv8 的红外图像行人检测算法，解决了传统 YOLOv8 算法在红外图像行人检测中存在的问题。本算法成功实现轻量化目标，助力在硬件资源平台有限的设备中的应用，同时兼顾了检测精度的要求。本文通过替换主干网络，提升模型在总体上对行人的特征提取能力的同时降低模型体积。其次，通过引入 Triplet 注意力机制，消除了通道与权重之间的间接对应。再次，基于可变形卷积重新构建了 C2f 模块，增强对不同形状的目标提取能力。最后在 Detect 引入一种新型卷积并替换部分原有卷积，进一步实现轻量化。实验表明，与现有模型相比，改进后的 RT-YOLOv8 算法模型体积大小降低的同时很好兼顾检测精度，满足了实时性与准确性平衡的需求，具有更为实用的价值。

第5章 基于红外-可见光图像跨模态融合的行人检测算法研究

5.1 引言

在昼间高照度道路场景下，现有基于可见光成像的行人检测算法已能达成预期应用标准。然而，夜间低照度道路环境中的行人检测才是交通安全领域的核心挑战。受限于人类视觉系统的生理特性，传统可见光成像在弱光照条件下存在显著的目标辨识缺陷，而基于红外光谱的检测技术为此提供了有效的解决方案。需要指出的是，夜间道路环境并非完全缺乏光照条件，诸如道路照明设施、自然月光等光源均会产生微弱的可见光辐射。但实验数据表明，此类低照度条件下的可见光强度（通常低于 10 lux）不足以支撑可靠的目标特征提取。在可见光/近红外混合光谱场景中，由于可见光通道信噪比(SNR<5dB)的急剧衰减，导致目标边缘特征模糊化(PSNR<20dB)、纹理信息熵值降低(≤1.2bits/pixel)等特征退化现象，使得传统卷积神经网络的层级特征表达能力显著下降。这种典型的光照受限场景下的多模态特征融合与鲁棒性检测问题，已成为计算机视觉领域亟待解决的技术难题。

红外热成像技术在夜间光照条件较差的情况下，可以与可见光图像进行取其之长补己之短的作用。通常采用特征融合网络分别提取特征后进行检测，以此来获得更多可用信息。图 5-1 展示了两种夜间具备光照条件下的可见光与红外图像。从图 5-1（a）中可以看出，在夜间光照较好的情况下，可以清晰地识别到附近的行人，但是由于在光照强度范围内，人类视觉视锥细胞敏感性会随着背景光强度的增强而成比例地下降。当与对面会车时远光灯或远处的直射光增强时，视锥细胞敏感性降低，会出现瞬间致盲的现象，持续时间大概有几秒，然而就这短暂的时间内驾驶人员对周围行人观察能力大大下降，可能会造成交通事故的发生。图 5-1（b）展示了夜间光照条件较差的情况下的可见光与红外图像，可以看出在光照昏暗的条件下，远处的行人几乎无法判断其大致位置，红外图像由于其成像特点可以很好弥补这一不足，然而红外图像由于采样设备的影响无法做到识别远距

离行人目标的缺陷，会造成目标漏检或误检的情况出现。所以基于跨模态融合的夜间场景行人检测将作为本章研究内容。

（a）夜间光照良好　　（b）夜间光照差

图 5-1 夜间双模态图像

5.2 多模态的图像融合方法

按照多模态图像融合的不同表征层次[65]可以将图像融合分为：像素级融合、特征级融合、决策级融合。

像素级融合属于低层次的融合，其技术特征体现在对原始数据流的直接操作。融合后的图像信息质量得到大幅提升，在遥感技术、医学影像中广泛应用。其主要通过三个流程：图像变换、变换系数融合、逆变换。

决策级融合属于高层次的融合，其在多个传感器独立的情况下，将多个传感器识别结果进行融合做出最终决策。其具有很好的纠错能力，通过适当的融合，消除单个传感器带来的负面效果。

特征级融合属于中间层次的融合，其对图像进行特征抽取，将图像的边缘等信息综合处理，减少了原始数据的处理量同时提高了系统的实时性。特征级融合分为前期、中期和后期融合。其中，前期、中期和后期融合示意图分别如图 5-2、图 5-3 和图 5-4 所示。

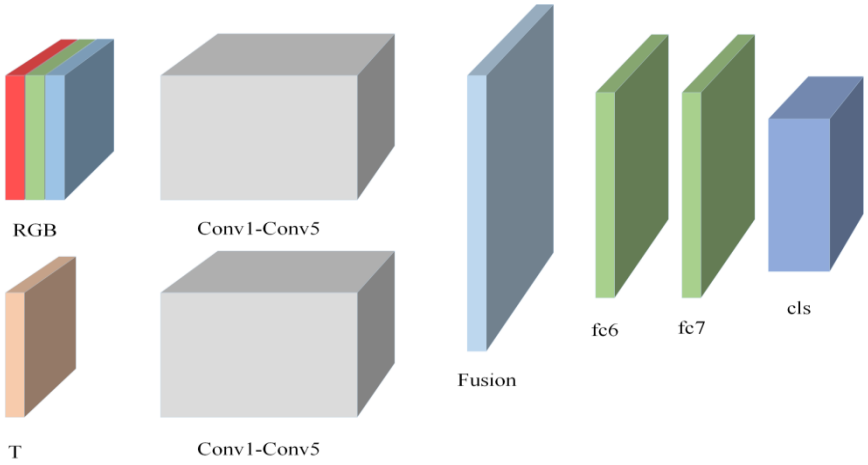


图 5-2 前期融合

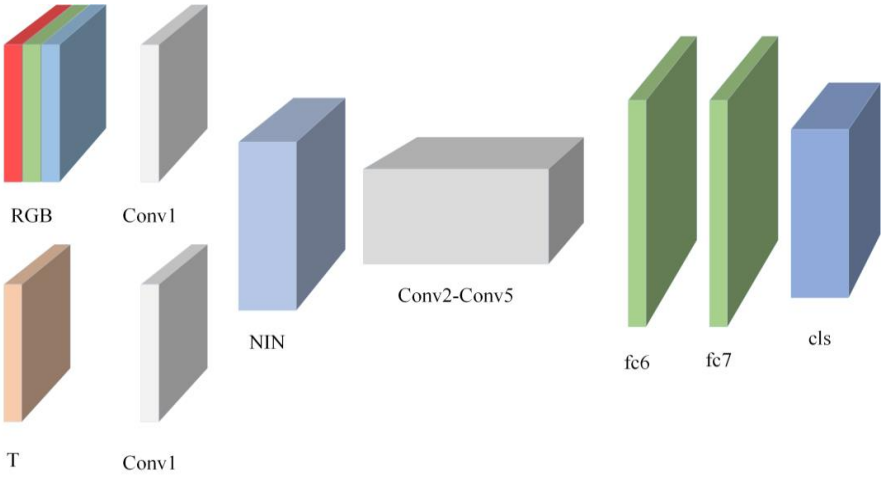


图 5-3 中期融合

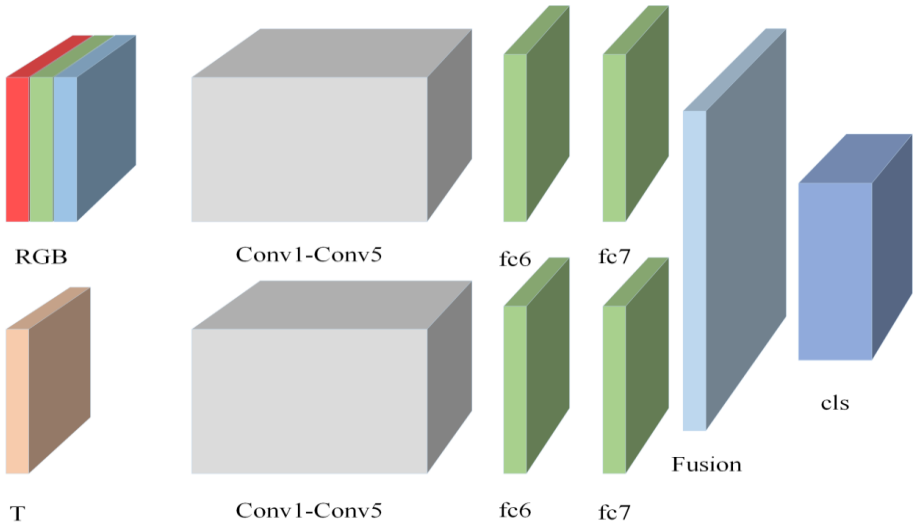


图 5-4 后期融合

多模态数据融合方法根据特征整合阶段可分为三种典型模式：（1）早期融合在数据输入后采用卷积层进行初步特征编码，并借助 NIN(Network-in-Network)

架构构建跨模态特征交互模块，随后进行深度特征学习与分类决策；（2）中期融合机制通过在双分支特征提取网络中引入 NIN 级联结构实现模态间特征关联，经特征组合后完成后续处理；（3）后期融合方案则是在各模态独立完成全连接层特征表示后，于分类层前实施最终的特征整合。。

根据已有跨模态行人检测的研究，在只有部分可见光的情况下，单一地通过采用可见光图像或者红外图像来进行行人检测，其检测精度往往较低，所以通过多种模态的图像融合后往往能够较单一场景的模态的行人检测各种性能会有所提升。表 5-1 是采用不同融合方式在本文建立的跨模态行人检测数据集上的对比实验。加粗数据表示当前指标最优值，底部加横线数据表示当前指标次优值。通过该表可以发现在单一模态可见光的场景情况下，其平均精度与模型参数量分别为 81.10%、11.02M，而单一模态红外图像的场景情况下，其平均精度和模型参数量分别为 78.10%、10.14M。后续跨模态融合对比实验各项检测结果均高于单一模态场景。在各种融合方式对比实验中，中期融合平均精度和模型参数量分别为 82.90%、9.29M，其结果在其他类型的融合方式中表现优异。所以在本文中采用中期融合进行夜间场景下的跨模态行人检测实验。

表 5-1 不同融合方式性能对比

融合方式	Precision (%)	Recall (%)	mAP@0.5 (%)	Parameter (M)
visible light	81.0	71.5	81.1	11.02
Infrared	79.8	72.4	78.1	<u>10.14</u>
像素融合	<u>82.3</u>	<u>73.6</u>	<u>81.7</u>	16.23
决策融合	81.9	73.0	81.5	12.54
特征融合-前期融合	81.3	72.7	81.2	11.47
特征融合-中期融合	**82.9**	**74.1**	**81.8**	**9.29**
特征融合-后期融合	78.0	71.1	77.2	15.22

为进一步提高在夜间具备可见光的场景下，有效提取红外-可见光图像的特征信息，并且在检测精度提升的同时保证模型参数量的情况下，本文采用双分支结构的中期融合网络分别提取红外-可见光图像的特征信息，然后在中期对双分支网络所提取的特征信息进行融合。

5.3 基于跨模态的行人检测网络设计

5.3.1 跨模态特征提取网络结构设计

为了将红外-可见光图像的特征信息可以更好地利用，本章设计了一种基于双分支并行特征提取架构的跨模态融合网络，由双分支来分别提取红外-可见光图像的行人特征信息，然后将双分支所提取的特征信息进行中期融合，以此来提升模型对图像信息的利用率，从而提高模型行人检测精度。

由于第四章提出的主干网络 RepGhostNet 使用 timm 库，其结构与本章所采用的中期融合方法相冲突，所以特征提取分支仍然采用原 YOLOv8 主干网络，并在其基础上进行模块改进，最后在 SPPF 模块后添加第四章所提出的 Triplet Attention 进行嵌入式设计。本章所设计的双分支特征提取网络结构如图 5-5 所示。

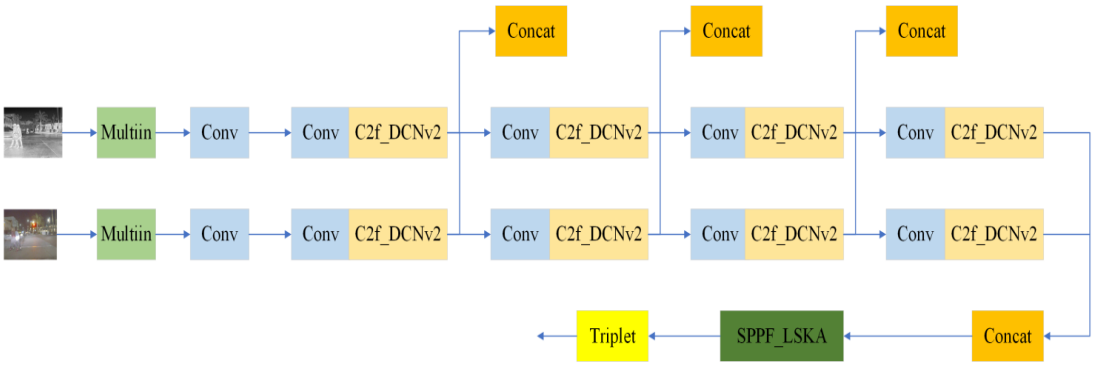


图 5-5 双分支特征提取网络结构图

5.3.2 LSKA 嵌入式设计

为了使模型聚焦于红外-可见光图像中行人的相关特征，从而进一步提取目标特征信息，增强对跨尺度特征的感知能力，在主干网络 SPPF 模块引入一种大可分离核注意力(Large Separable Kernel Attention, LSKA)。LSKA 通过对大核卷积进行分解操作，如图 5-6（a）所示，将大核卷积分解为通道卷积、深度扩张卷积核深度卷积通过利用大且可分离卷积核来捕捉图像是广泛上下文信息，生成注意力图，通过注意力图加权原始特征，从而使得网络模型针对行人特征信息。

该注意力机制首先将一个 k×k 的大核卷积分解成(2d-1)×(2d-1)的通道卷积、k/d×k/d 的深度扩张卷积和 1×1 的深度卷积。最后将 2D 卷积分离成 1D 卷积，

其过程如下：首先将 2D 的 DW 卷积核与 DW-D 卷积继续分解为 1D 的水平卷积核与垂直卷积核，然后将分解后的卷积核按照顺序串联。其结构示意图如图 5-6（b）所示。

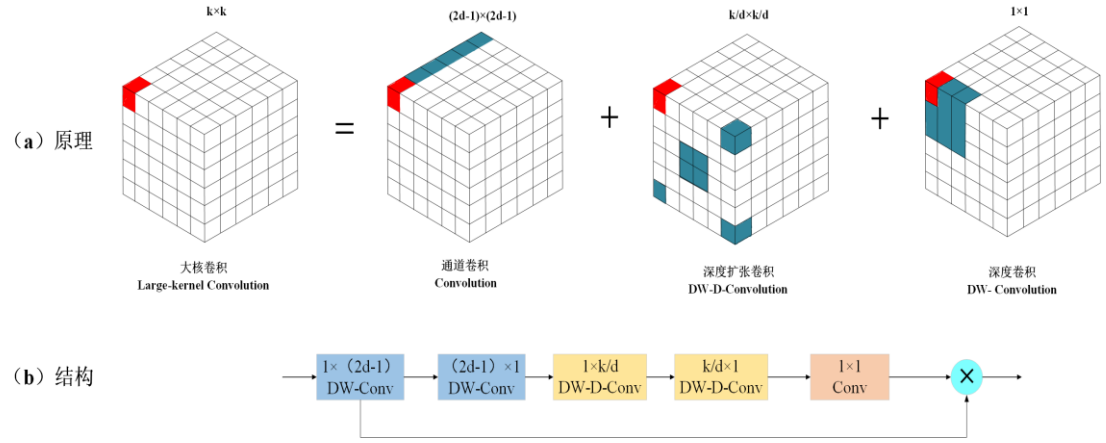


图 5-6 LSKA 原理及结构示意图

对原始主干网络 SPPF 模块进行修改，其 SPPF-LSKA 模块结构图如图 5-7 所示。首先特征信息通过一个卷积层后再连续通过三个相同的最大池化层，然后上述输出通过 Concat 模块连接在一起输入到 LSKA 模块中进行处理，最后通过一个普通卷积输出特征信息。

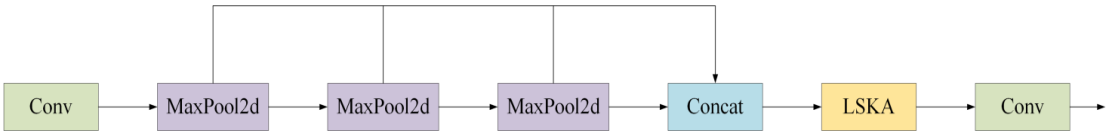


图 5-7 SPPF-LSKA 结构图

5.3.3 损失函数的改进

边界框回归(Bounding Box Regression, BBR)在目标检测中被广泛应用于优化模型性能中，由于损失函数的选取对模型的后续训练及收敛速度至关重要，当模型收敛速度越快的情况下，其可以获得更好的性能。同时，损失函数也决定着模型的泛化能力，即模型对其他未经过训练数据的处理能力。

通过引入 MPDIoU 损失函数，基于现有 *IoU-based* 损失和 $l_n - norm$ 损失的优缺点的基础上，采用一种基于最小点距离的 IoU 损失，即 MPDIoU。其解决了预测框与真实框纵横比相同使宽度和高度值却不同情况下的问题，加快了模型的收敛速度。MPDIoU_loss 计算因素如图 5-8 所示。

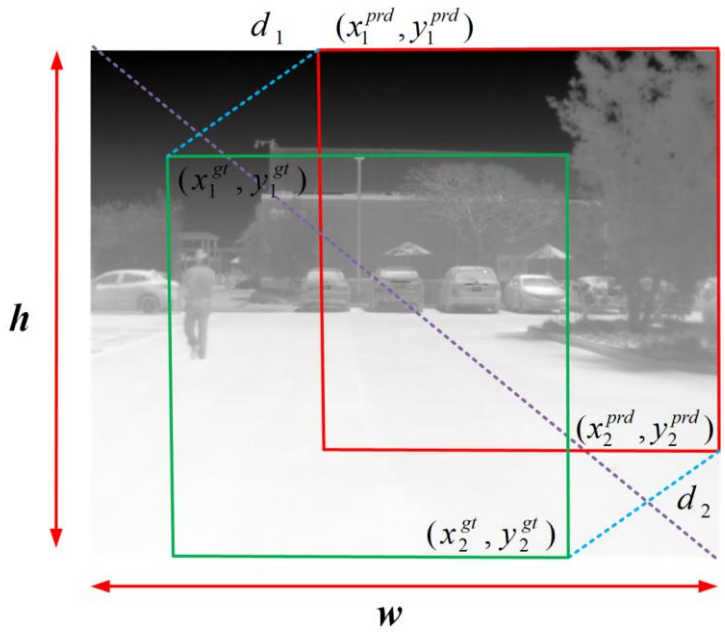


图 5-8 MPDIoU_loss 计算因素图示

MPDIoU 的计算过程如下所示。

$$d_1^2 = (x_1^{prd} - x_1^{gt})^2 + (y_1^{prd} - y_1^{gt})^2 \tag{5-1}$$

$$d_2^2 = (x_2^{prd} - x_2^{gt})^2 + (y_2^{prd} - y_2^{gt})^2 \tag{5-2}$$

$$MPDIoU = IoU - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \tag{5-3}$$

$$|C| = (\max(x_2^{gt}, x_2^{prd}) - \min(x_1^{gt}, x_1^{prd}))(\max(y_2^{gt}, y_2^{prd}) - \min(y_1^{gt}, y_1^{prd})) \tag{5-4}$$

$$x_c^{gt} = \frac{(x_2^{gt} + x_1^{gt})}{2}, y_c^{gt} = \frac{(y_2^{gt} + y_1^{gt})}{2} \tag{5-5}$$

$$x_c^{prd} = \frac{(x_2^{prd} + x_1^{prd})}{2}, y_c^{prd} = \frac{(y_2^{prd} + y_1^{prd})}{2} \tag{5-6}$$

$$w_{gt} = x_2^{gt} - x_1^{gt}, h_{gt} = y_2^{gt} - y_1^{gt} \tag{5-7}$$

$$w_{prd} = x_2^{prd} - x_1^{prd}, h_{prd} = y_2^{prd} - y_1^{prd} \tag{5-8}$$

其中、(x_1^{gt}, y_1^{gt})、(x_2^{gt}, y_2^{gt})分别代表真实框的左上角和右下角坐标，(x_1^{prd}, y_1^{prd})、(x_2^{prd}, y_2^{prd})分别代表预测框的左上角和右下角坐标，w、h 分别代表输入图像的宽和高，$|C|$代表预测框和真实框的最小包围矩形面积，(x_c^{gt}, y_c^{gt})、(x_c^{prd}, y_c^{prd})分别代表真实框和预测框的中心点坐标，w_{gt}、h_{gt} 分别代表真实框的宽和高，w_{prd}、h_{prd} 分别代表预测框的宽和高。MPDIoU_loss 定义公式如式(5-9)所示。

$$MPDIoU\_loss = 1 - MPDIoU = 1 - IoU + \frac{d_1^2}{w^2 + h^2} + \frac{d_2^2}{w^2 + h^2} \tag{5-9}$$

5.4 实验结果与分析

5.4.1 实验配置与参数

为了验证本章所提出跨模态行人检测算法的有效性，本章采用第三章所建立的红外-可见光行人数据集上进行实验，改进网络在红外-可见光行人数据集上训练的损失函数变化如图 5-9 所示，实验结果显示当迭代次数达到 200 轮次时，损失函数曲线进入平稳状态，表明模型已达到优化稳定阶段，本章最终选定 200 轮次为模型训练周期。由于本章针对夜间具备一定可见光场景下行人检测，这种场景下具有一定数量的骑行者，所以为了让本章实验数据一目了然，实验评价指标均针对所有类（包括步行者、骑行者）。

实验在系统 ubuntu22.04 系统环境下进行，硬件设备为显卡 8G 显存的 NVIDIA GeForce RTX1070，软件环境为 python3.8 使用 torch1.13.1 框架，设置超参数 learning rate 为 0.01，batch size 为 16。

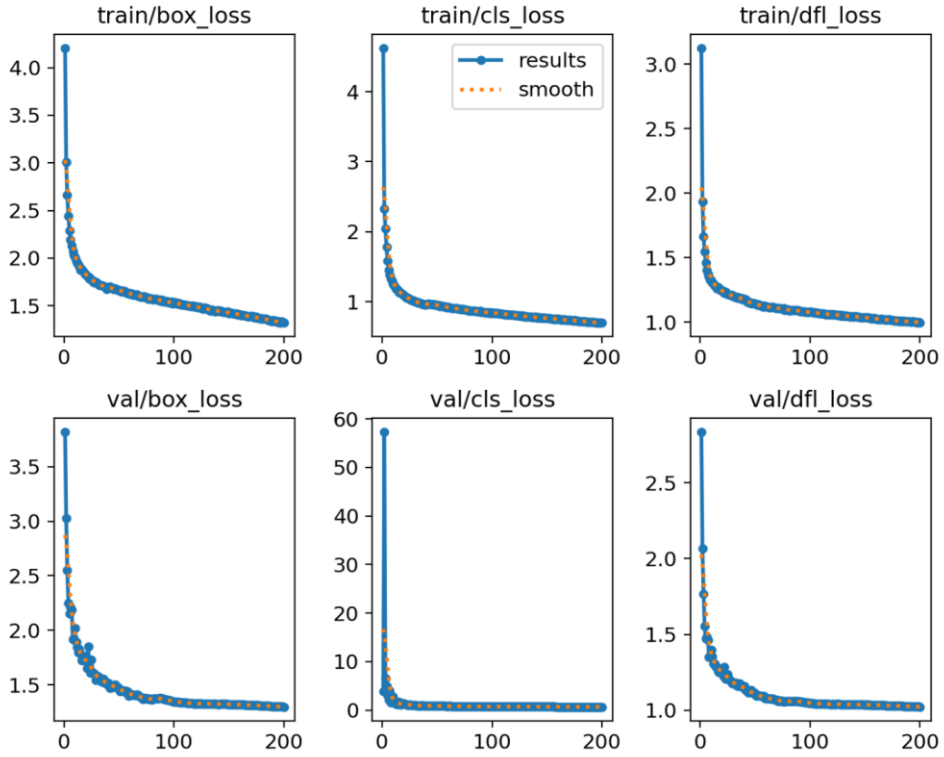


图 5-9 损失函数变化图

5.4.2 消融实验

为进一步验证各改进模块之间的组合性能设计消融实验，探究方法的有效性，本章在基于中期融合的 YOLOv8 算法的基础上进行实验，本章每组实验均使用相同的超参数设置和训练技巧。本章实验在第三章建立的红外-可见光图像行人数据集上进行，消融实验结果如表 5-2 所示，其中，“√”表示使用改进方法，“-”表示未使用改进方法。加粗数据表示当前指标最优值，底部加横线数据表示当前指标次优值。

表 5-2 消融实验结果

SPPF_ LSKA	C2f_ DCNv2	Triplet	Detect _PC	MP_ DIoU	Precision (%)	Recall (%)	mAP@0.5 (%)	Parameter (M)
-	-	-	-	-	82.1	73.2	80.4	11.47
√	-	-	-	-	82.4	74.9	81.8	10.33
-	√	-	-	-	80.9	74.5	81.9	9.88
-	-	√	-	-	82.2	75.2	82.0	9.29
-	-	-	√	-	81.2	74.9	81.5	**7.06**
-	-	-	-	√	82.6	74.5	82.0	9.29
√	√	-	-	-	83.0	75.6	82.7	10.92
	√	√	-	-	**84.7**	75.2	83.5	9.88
√	-	-	√	-	79.3	76.4	82.2	10.77
√	-	√	-	-	<u>83.1</u>	75.1	82.7	10.34
√	√	√	-	-	82.9	<u>77.1</u>	<u>83.7</u>	10.92
-	√	√	√	-	81.9	73.9	82.3	8.70
√	√	√	√	-	82.7	75.9	83.0	8.86
-	√	√	√	√	82.3	76.5	82.6	8.76
√	√	√	√	-	82.3	76.1	83.6	8.72
√	√	√	√	√	82.5	**77.4**	**84.6**	<u>8.69</u>

由表 5-2 消融实验结果表可知，在针对基于中期融合的 YOLOv8 算法的基础上，通过将 LSKA 嵌入到 SPPF 模块中，平均精测精度为 81.80%；通过引入第四章提出的 C2f_DCNv2 模块、Triplet 注意力机制和 Detect_PC 模块，其平均检测精度分别为 81.90%、82.00%和 81.50%；通过改进损失函数，其平均检测精度为 82.00%；使得模型整体检测精度较之原始算法平均精度有所提升，并且模型参数量较之也有所降低。通过最后的消融实验结果可知，在基于跨模态的行人检测算法上，本章算法精测平均精度和召回率均有明显的提升，均提高 4.2%，并且算法模型参数量在原始算法的基础上降低了 24.29%。

5.4.3 主流算法对比实验

为验证融合图像的必要性和可行性，评估本文算法的整体性能设计对比实验，其中涵盖了红外图像、可见光图像和双模态图像分别进行对比，对比实验在本文建立的红外-可见光图像数据集上进行。模型检测性能对比结果如表 5-3 所示。加粗数据表示当前指标最优值，底部加横线数据表示当前指标次优值。

表 5-3 红外、可见光、融合图像对比实验

Input	Algorithm	Precision (%)	Recall (%)	mAP@0.5 (%)
infrared	YOLOv3	78.2	73.0	77.9
	YOLOv5	81.0	75.3	82.6
	YOLOv7	79.6	71.9	80.2
	SSD	74.9	69.1	73.1
	RetinaNet	72.9	71.5	75.4
	YOLOv8	<u>81.4</u>	**79.6**	83.1
	YOLOv9	81.3	77.1	<u>83.2</u>
	YOLOv10	79.5	76.2	81.2
	YOLOv11	79.0	71.4	81.9
visible	YOLOv3	77.2	71.9	75.4
	YOLOv5	80.4	72.7	78.2
	YOLOv7	80.0	74.6	71.0
	SSD	75.3	70.1	77.4
	RetinaNet	79.8	69.4	75.3
	YOLOv8	80.3	75.5	82.4
	YOLOv9	79.8	75.1	83.0
	YOLOv10	79.0	70.2	82.1
	YOLOv11	77.9	71.0	82.0
fusion	MMTOD	80.2	73.5	82.7
	CMDet	70.9	69.3	72.6
	CFT	81.0	72.9	81.5
	RISNet	79.9	75.1	80.3
	Ours	**82.5**	<u>77.4</u>	**84.6**

在数据集上实验由表 5-3 可知：在单一模态图像与融合后图像检测效果对比中，YOLOv8 算法表现较为优异，精确度为 81.40%，在三组对比实验中表现次优，召回率为 79.60%，在三组对比实验中表现最优，YOLOv9 平均精度为 83.20%，在三组对比实验中表现次优。由于夜间可见光图像分辨率较低，所以导致对比实验中在可见光图像上的实验效果较差。在融合后图像上，采用了一些新型跨模态目标检测算法(MMTOD[68]、CMDet[69]、CFT[70]、RISNet[71])以及本章提出的算法。本章算法精确度、召回率和平均精度分别为 82.50%、77.40%和 84.60%，较之红外图像上的原算法精确度、召回率和平均精度分别提高 1.10%、-2.20%和 1.50%，较之可见光图像上的原算法精确度、召回率和平均精度分别提高 2.20%、1.90%

和 2.20%，提升较为明显。在跨模态图像上进行的对比实验中优势十分明显，较之其他算法有较大幅度的精度提升。另外从图 5-10 和图 5-11 中可以看出，从精确率和召回率的定义来理解，中期融合后改进算法精确率和召回率相较于原算法检测单一模态指标较高，从中可以得出以下结论，本章的模型在夜间跨模态行人检测算法中有较好地表现。

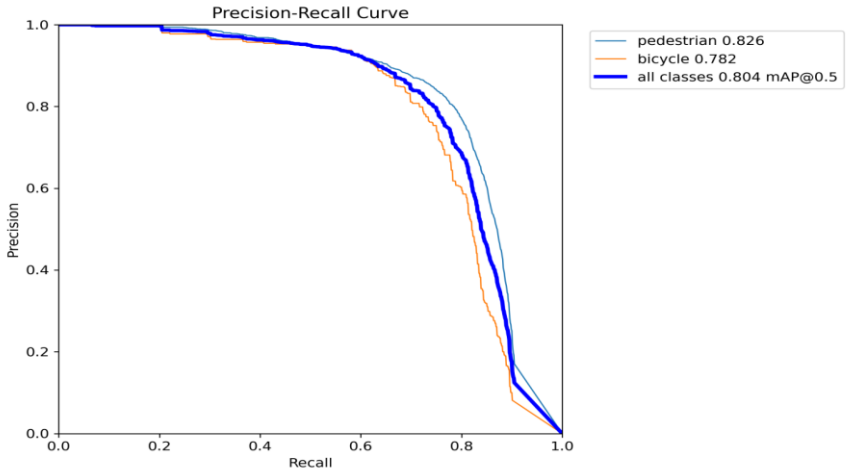


图 5-10 中期融合后原算法 Precision-Recall Curve 曲线图

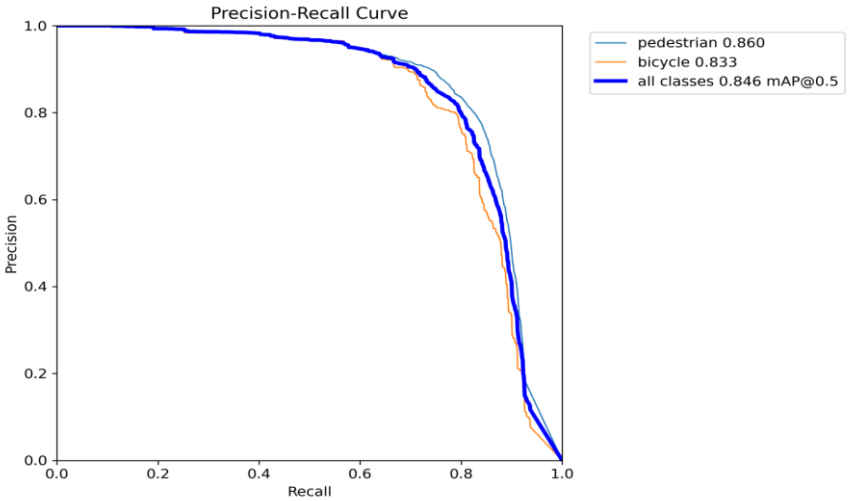


图 5-11 中期融合后改进算法 Precision-Recall Curve 曲线图

为了进一步证明改进算法的可靠性，在公开红外行人检测数据集 KAIST 上进行多种算法(GAFF[72]、IAF R-CNN[73]、IATDNN+IASS[74])和本章改进 YOLOv8 对比实验。

GAFF 算法通过提出一种自适应的特征融合方法即引导注意力特征融合，通过两种模态的注意力之间的强化，使得模型学习到自适应加权和融合。IAF R-CNN 通过引入置信度融合体系来提高精测精度。IATDNN+IASS 通过多个子网络

来检测目标，然后对检测结果进行融合后生成最终结果。对比实验结果如表 5-4，其中加粗数据表示当前指标最优值，底部加横线数据表示当前指标次优值。

表 5-4 KAIST 数据集测试结果

Algorithm	Precision (%)	Recall (%)	mAP@0.5 (%)
GAFF	75.9	80.1	81.7
IAF R-CNN	79.3	81.9	<u>87.3</u>
IATDNN+IASS	<u>85.1</u>	**85.3**	83.6
Ours	**87.7**	<u>85.1</u>	**88.5**

在公开红外行人检测数据集上实验得出：在其他三种对比算法中，IATDNN+IAS 精确度数据最优为 85.10%，同样其召回率也是最优的，为 85.30%，而在平均精度评价指标上表现最优的为 IAF R-CNN 为 87.30%，本文算法精确度为 87.70%，在对比算法中表现最优；召回率为 85.10%，在对比算法中表现次优；平均精度为 88.50%，在对比算法中表现最优。从上述数据可以得知，本文算法虽然召回率为次优，但是其他两项性能指标表现最优，验证了本章实验的有效性，达到了本章预期的效果。

5.4.4 实验结果可视化分析

为了更加直观地比较本章所提出的基于跨模态行人检测算法的有效性，分别采用单模态 YOLOv8 算法对红外图像和可见光图像进行检测，检测效果图如图 5-12、图 5-13 所示。

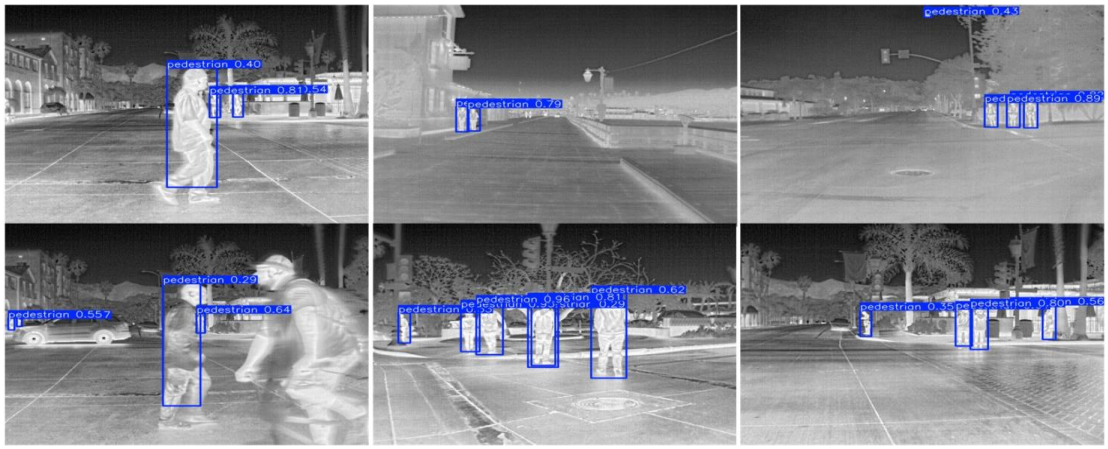


图 5-12 红外模态检测效果图

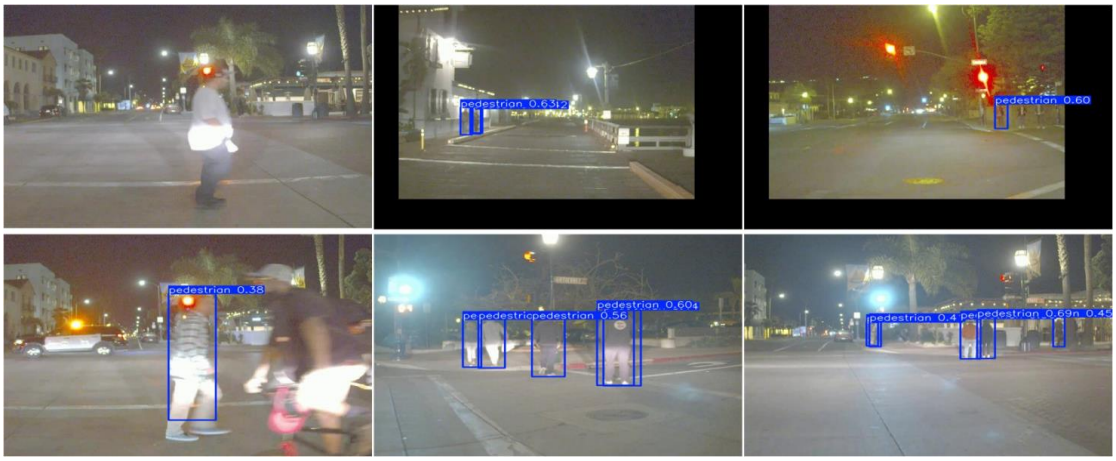


图 5-13 可见光模态检测效果图

图 5-14 跨模态检测效果图

从图 5-12、图 5-13 中可以发现在单一模态的情况下，行人检测存在部分漏检。红外模态下的检测图，在图 5-12 第二行左边第一张图片结果显示，由于目标无法与周围场景形成对比，会导致检测精度较低甚至无法检测的情况。可见光模态下的检测图，在图 5-13 第一行左边第一张图片结果显示，由于夜晚灯光的反射导致目标产生眩晕的情况。部分较远场景下的小目标无法很好地识别，导致存在漏检的情况。如图 5-14 展示出了本章跨模态融合算法检测效果图，从图中可以发现本章算法在结合双模态图像特征后，能够准确地检测到行人目标，显示出其在纠正基准算法误检漏检的错误以及提高检测精度方面的显著优势。这些结果充分地体现了本章算法在目标检测任务中的优异性能。

5.5 本章小结

本章首先介绍跨模态图像多种常见融合方法，然后基于中期融合的方法上设

计一种双分支特征提取网络，分别提取红外图像和可见光图像的特征后通过中期融合进行特征融合。同时为了提高特征提取的效率，在双分支主干网络 SPPF 模块中引入 LSKA。经过消融实验筛选出的第四章提出的 C2f_DCNv2 模块、Triplet 注意力机制和 Detect_PC 模块将其融合于跨模态行人检测网络中。针对原始网络损失函数的不足，将其替换为 MPDIoU 损失函数，解决原始损失函数预测框与真实框具有相同的纵横比，但宽度和高度值不同情况下的问题，并且加快了模型的收敛速度。最后为验证本章所提出算法的有效性，设计消融实验和对比实验，实验数据对比分析显示，本文设计的跨模态融合算法在行人检测任务重的识别率较传统单一模态方法提升显著，充分证明了跨模态协同感知的有效性。

第6章 总结与展望

6.1 全文总结

由于科技的不断进步，无人驾驶技术也不断随之提升。然而在以人为本的前提下，如何能够在各种场景下高精度的识别并划分行人目标则显得尤为重要了，但是基于夜间道路场景的行人检测研究仍存在显著的技术突破空间，其往往可以应用在可视条件差的场景中，现在大部分车载采集设备是针对可见光的，红外成像技术可以很好地弥补这一劣势。然而受限于红外设备造价高昂等因素，目前仍然存在应用局限性。尽管如此，随着无人驾驶技术的产业化进程加速，全天候环境感知能力已成为核心需求，这直接导致对于夜间道路场景的行人检测需求则会日益增加。

目标检测不仅要求算法识别待测目标的位置，还需要对其识别分类。如何提高夜间道路场景的行人检测精度是本文研究的主要内容，为了促进夜间行人检测技术的发展，本文研究基于夜间道路场景下的行人检测算法，研究内容包括三个方面：

（1）针对夜间红外图像进行预处理，并划分建立本文所需的两个夜间红外数据集。针对红外图像目标与背景对比度低的特性，使用 DDE 算法对其进行局部特征增强，包括提高目标的对比度和图像的清晰度；人工制作标签并划分为红外图像数据集和红外-可见光图像跨模态数据集，最后自动划分训练集和验证集。

（2）针对红外场景下夜间道路场景，设计了一种兼顾实时性与准确性的行人检测算法 RT_YOLOv8。通过采用 RepGhostNet 框架作为主干网络来达到实时性的目的，为了提高行人检测的准确性；在主干网络末端引入 Triplet Attention 来增加模型的表征能力；同时对部分 C2f 模块进行改进，通过引入可变形卷积 DCNv2 使其更好地适应多种形态的行人目标；最后通过对 Detect 模块进行重构，通过将部分常规卷积替换为 PConv，使其模型参数量进一步得到减少。最后通过实验验证了所提出算法的有效性，达到了预期目标。

（3）针对低照度复合场景（如月光、路灯环境），设计了一种基于跨模态的行人检测网络。首先通过采用双分支网络分别对红外-可见光图像进行特征提

取，然后采用中期融合的方法对所提取的特征信息进行融合，在特征提取阶段，通过在 SPPF 模块中嵌入大核注意力机制(LSKA)，增强对跨尺度特征的感知能力，再对损失函数的存在的问题对其进行替换，优化模型对其他未经过训练数据的处理能力。最后通过引入第四章提出的 C2f_DCNv2 模块、Triplet 注意力机制和 Detect_PC 模块，将其应用在跨模态行人检测网络中。通过后续实验进一步验证，实验数据结果表明本章提出算法能够很好地进行跨模态图像行人检测。

6.2 未来展望

本文针对 YOLOv8 行人检测算法在可见光不足的夜间道路场景优化其检测效果，主要提出了两种夜间道路场景行人检测改进算法，两种算法分别从不同角度去分析具体情况在提高精度的同时兼顾模型检测的实时性，但仍然存在一定改进空间。在未来工作中，可以从以下几个方面进行深入研究：

（1）拓展数据集，增加数据集的多样性，从而面向不同形态的行人。由于行人作为一种非静止目标，人体一旦发生运动，其形态特征会发生较为明显的变化，如下蹲时其形态与直立形态具有较大差异。然而本文对于多形态的行人目标还未具有足够的泛化性。在后续的研究中，可以通过对实验数据集进行扩充，增加其多形态行人的模型训练，从而提升夜间场景下的行人检测的泛化性。

（2）解决行人遮挡问题。在本文数据集中虽然存在部分场景下行人遮挡图像，可以解决部分行人遮挡问题，然而在现实生活中，遮挡情况随时会发生，存在各种不同条件下的行人遮挡问题，其带来最直接的影响就是行人目标的不完全。因此，如何更有效地解决行人遮挡问题，降低行人目标的误检率，将成为后续研究的重要方向。

（3）解决非行人热源问题。在复杂的场景中，红外图像采集会受其他热源干扰，如机动车发动机、夏季道路。它们在红外图像中形成高亮区域，严重影响了行人目标的识别，从而对检测效果造成影响。因此如何有效地降低红外图像中其他非行人热源的干扰，提高行人检测的准确性具有较大的应用潜力。

参考文献

[1] Liu L, Ouyang W, Wang X, et al. Deep learning for generic object detection: A survey[J]. International journal of computer vision, 2020, 128: 261-318.

[2] Henini M, Razeghi M. Handbook of infrared detection technologies[M]. Amsterdam: Elsevier, 2002.

[3] Xiong Y A N, Peng J X, Ding M Y, et al. An extended track-before-detect algorithm for infrared target detection[J]. IEEE transactions on aerospace and electronic systems, 1997, 33(3): 1087-1092.

[4] Andraši P, Radišić T, Muštra M, et al. Night-time detection of uavs using thermal infrared camera[J]. Transportation Research Procedia, 2017, 28: 183-190.

[5] Heo D, Lee E, Chul Ko B. Pedestrian detection at night using deep neural networks and saliency maps[J]. Journal of Imaging Science and Technology, 2017, 61(6): 604031-604039.

[6] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[7] Ren S, 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, 2015, 39(6) : 1137-1149.

[8] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[9] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[10] Ren S, He K, Girshick R, et al. Faster r-cnn: towards real-time object detection with region proposal networks[J]. Advances in Neural Information Processing Systems, 2015, 28.

[11] HE K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE International Conference on Computer Vision, 2017: 2961-2969.

[12] Gavrilescu R, Zet C, Fosalau C, et al. Faster R-CNN: an approach to real time object detection[C]//Proc of International Conference and Exposition on Electrical and Power Engineering, 2018: 165-168.

[13] CAI Z,Vasconcelos N.Cascade R-CNN:delving into high quality object detection [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6154-6162.

[14] HE Kaiming, Gkioxari Georgia, Dollar Piotr, et al. Mask R-CNN[J]. IEEE

Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 318-327.

[15] REN S, 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(06): 1137-1149.

[16] Bochkovskiy A, WANG C Y, LIAO H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv: 2004.10934, 2020.

[17] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. arXiv preprint arXiv: 1804.02767, 2018.

[18] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7263-7271.

[19] LIU W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector[C]// European Conference on Computer Vision, 2016: 21-37.

[20] LIU W, ANGUELOV D, ERHAND, et al. SSD: single shot multibox detector [C]//Proc of IEEE Conference on Computer Vision and Pattern Recognition. Berlin: Springer, 2016: 21-37.

[21] Viola P, Jones M. Rapid Object Detection Using a Boosted Cascade of Simple Features[C]//IEEE Computer Society, 2001, 57(2): 34-47.

[22] Zhang L, Wu B, Nevatia R. Pedestrian Detection in Infrared Images Based on Local Shape Features[C]//IEEE Computer Society on Computer Vision and Pattern Recognition, 2007: 1-8.

[23] Olmeda D, Escalera A D L, Armingol J M. Contrast Invariant Features for Human Detection in Far Infrared Images[J]. 2010, 5(3): 117-122.

[24] Wang X, Han X, Yan S. A HOG-LBP Human Detector with Partial Occlusion Handling[C]//Proceedings of International Conference on Computer Vision. Kyoto: IEEE Computer Society, 2009: 32-39.

[25] Mi R J, Kwak J Y, Son J E, et al. Fast Pedestrian Detection Using a Night Vision System for Safety Driving[C]//International Conference on Computer Graphics, Imaging and Visualization. IEEE, 2014: 69- 72.

[26] 刘峰, 王思博, 王向军.多特征级联的低能见度环境红外行人检测方法[J]. 红外与激光工程, 2018, 47(6): 137-144.

[27] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20: 273-297.

[28] 王晓红, 陈哲奇. 基于 YOLOv5 算法的红外图像行人检测研究[J]. 激光与红

外, 2023,53(01): 57-63.

[29]代少升, 刘科生, 黄炼, 等. 基于视觉 Transformer 和双解码器的红外小目标检测方法[J]. 红外技术, 2024, 46(09): 1070-1080.

[30]刘学, 李范鸣, 刘士建. 改进的 SSD 红外图像行人检测算法[J]. 电光与控制, 2020, 27(01): 42-46+59.

[31]HWANG S,PARK J,KIM N ,et al. Multispectral pedestrian detection: benchmark dataset and baseline[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition,2015: 1037-1045.

[32]KONIG D, ADAM M, JARVERS C, et al. Fully convolutional region proposal networks for multispectral person detection[C]//Proceedings of the 2017 IEEE Conference on Computer Vision & Pattern Recognition Workshops,017: 243-250.

[33]GUAN D Y, CAO Y, YANG J, et al. Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection [J]. Information Fusion, 2019, 50: 148-157.

[34]LI C,SONG D,TONG R, et al. Illumination-aware faster R-CNN for robust multispectral pedestrian detection[J]. Pattern Recognition, 2019, 85: 161-171.

[35]LI C, SONG D, TONG R, et al. Multispectral pedestrian detection via simultaneous detection and segmentation[C]//Proceedings of the British Machine Vision Conference, 2018: 225-234.

[36]ZHANG L, LIU Z, ZHANG S, et al. Cross-modality interactive attention network for multispectral pedestrian detection[J]. Information Fusion, 2018,50: 20-29.

[37]ZHANG L, ZHU X, CHEN X, et al. Weakly aligned cross-modal learning for multispectral pedestrian detection[C]//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, 2019: 5126-5136.

[38]ZHOU K, CHEN L, CAO X. Improving multispectral pedestrian detection by addressing modality imbalance problems[C]//Proceedings of the 2020 IEEE Conference on European Conference on Computer Vision, 2020: 787-803.

[39]FU L, GU W, AI Y. Adaptive spatial pixel-level feature fusion network for multispectral pedestrian detection[J]. Infrared Physics & Technology,2019,99:1-16.

[40]DASGUPTA K, DAS A, DAS S, et al. Spatio-contextual deep network based multimodal pedestrian detection for autonomous driving[B/ OL].

[41]H.Bhujle"Weighted-average fusion method for multiband images,"2016

International Conference on Signal Processing and Communications (SPCOM), Bangalore, India, 2016, pp. 1-5, doi: 10.1109/SPCOM.2016.7746635.

[42] Zhao S, Chen X, Wang S, et al. A new method of remote sensing image decision-level fusion based on Support Vector Machine[J]. IEEE, 2003.

[43] 刘峰, 沈同圣, 郭少军, 等.基于特征级融合的多波段舰船目标识别方法[J]. 光谱学与光谱分析, 2017, 37(06): 1934-1940.

[44] Li H, Wu X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2018, 28(5): 2614-2623.

[45] 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.

[46] 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.

[47] 娄熙承, 冯鑫. 潜在低秩表示框架下基于卷积神经网络结合引导滤波的红外与可见光图像融合[J]. 光子学报, 2021, 50(03): 188-201.

[48] Liu D, Zhou D M, Nie R C, et al. Infrared and visible image fusion based on convolutional neural network model and saliency detection via hybrid 10-11 layer decomposition[J].Journal of electronic imaging, 2018,27(6): 063036.1-063036.10.

[49] 王娟, 柯聪, 刘敏, 等. 神经网络框架下的红外与可见光图像融合算法综述[J]. 激光杂志, 2020, 41(07): 7-12.

[50] Singh Y, Saini M. Impact and Performance Analysis of Various Activation Functions for Classification Problems[C]//2023 IEEE International Conference on Contemporary Computing and Communications (InC4). IEEE, 2023, 1: 1-7.

[51] Dombi J, Jónás T. The generalized sigmoid function and its connection with logical operators[J]. International Journal of Approximate Reasoning, 2022, 143: 121-138.

[52] De Ryck T, Lanthaler S, Mishra S. On the approximation of functions by tanh neural networks[J]. Neural Networks, 2021, 143: 732-750.

[53] Agarap A F. Deep learning using rectified linear units (relu)[J]. arXiv preprint arXiv: 1803.08375, 2018.

[54] Alkhatib R. Artificial Neural Network Activation Functions in Exact Analytical Form (Heaviside, ReLU, PReLU, ELU, SELU, ELiSH)[J]. Authorea Preprints, 2023.

[55]孙晓刚, 李云红. 红外热像仪测温技术发展综述[J]. 激光与红外, 2008, (02): 101-104.

[56]Fang Y, Yamada K, Ninomiya Y, et al. Comparison between infrared-image-based and visible-image- based approaches for pedestrian detection[C]// IEEE Intelligent Vehicles Symposium. IEEE, 2003: 505-510.

[57]王欣, 徐平平, 吴菲. 基于指数同态滤波耦合细节锐化规则的红外图像增强算法[J]. 电子测量与仪器学报, 2021, 35(10): 9-16.

[58]路皓翔, 刘振丙, 张静, 等. 结合多尺度循环卷积和多聚类空间的红外图像增强[J]. 电子学报, 2022, 50(02): 415-425.

[59]ZHOU J L, ZHANG Y Y, WANG J P.RDE -YOLOv7: an improved model based on YOLOv7 for better performance in detecting dragon fruits[J]. Agronomy, 2023, 13(4): 1042.

[60]MISRA D, NALAMADA T, ARASANIPALAI A U, et al. Rotate to attend: convolutional triplet attention module[C]//oceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), 2021: 3138-3147.

[61]DAI J, Qi H, Xiong Y, et al. Deformable Convolutional Networks[C/OL]//2017 IEEE International Conference on Computer Vision (ICCV). 2017: 764-773.

[62]Zhu X, Hu H, Lin S,et al. Deformable ConvNets V2: More Deformable, Better Results[C]//2019 IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 9: 9300-9308.

[63]WANG W H, DAI J f, CHEN Zhe, HUANG Z H, LI A Q. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions [C]. arXiv: 2211. 05778, 2022.

[64]Chen J, Kao S, He H, et al. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks[J]. arXiv preprint arXiv: 2303.03667, 2023.

[65]韩晓军. 数字图像处理技术与应用[M].北京: 电子工业出版社, 2017: 151-159.

[66]张泓国.红外与可见光图像融合算法研究[D]. 兰州交通大学, 2022.

[67]储彬彬, 庞璐璐, 漆德宁, 等. 基于多尺度分析的图像融合技术综述[J]. 航空电子技术, 2009, 40(03): 29-33+48.

[68]Devaguptapu C, Akolekar N, M Sharma M and N Balasubramanian V. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery[C].

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2019: 0-0.

[69] Sun Y, Cao B, Zhu P and Hu Q. Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(10): 6700-6713.

[70] Fang Q, Han D, Wang Zi. Cross-modality fusion transformer for multispectral object detection[J]. arXiv preprint arXiv:2111.00273, 2021.

[71] Wang Q, Chi Y, Shen T, Song J, Zhang Z and Zhu Y. Improving RGB-Infrared Object Detection by Reducing Cross-Modality Redundancy[J]. Remote Sensing, 2022, 14(9): 2020-2031.

[72] Zhang H, Fromont E, Lefèvre S, et al. Guided attentive feature fusion for multispectral pedestrian detection[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 72-80.

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

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

攻读学位期间所发表的学术论文

[1] 凌成, 许钢, 温福新.红外图像行人检测实时性与准确性的平衡研究[J].红外技术(CSCD), 2024-07-19.(录用)

[2] 温福新, 许钢, 凌成, 等.基于改进 YOLOv10n 的红外图像遮挡行人检测算法[J].重庆工商大学学报(自然科学版), 1-11[2025-05-16].

[3] 温福新, 许钢, 凌成, 等.基于改进 YOLOv5s 的红外图像行人检测算法[J].贵州大学学报(自然科学版), 2025-03. (录用)

致谢

行文至此，落笔为终。回首三载求学路，此刻都化为心头的万千感慨。在此，谨以最真挚的文字，向所有照亮这段旅程的人们致以最深切的谢意。

首先，我要衷心感谢我的导师许钢教授。感谢许老师三年以来的谆谆教诲和殷切指导。从选题方向的确立到实验设计的优化，从数据纠偏到论文成稿，您严谨的治学态度与渊博的专业学识始终为我指引方向。难忘您的指导，逐页批注论文时留下的红色批注；您不仅授我以“术”，更以身作则教会我如何以理性思辨与人文关怀面对学术与人生。

其次，我要感谢实验室的所有同学们，温福新、苏世林、王博，在这个研究旅程中，有了你们的支持与合作，我才得以顺利进行实验室工作。我们共同协作的时光充满了团队精神和合作意识，成为我学术道路上难忘的一部分。

再次，感谢我的父母在这二十多年来对我的养育之恩，感谢你们对我无条件的支持和鼓励，感谢你们给与我人生更多的选择和机会。

最后，谨以此文献给所有在时代浪潮中坚持学术理想的同行者。当我们把论文写在祖国大地上时，那些在寂寞中沉淀的思考，终将在某个清晨绽放出改变世界的力量。