DOI: 10.13382/j.jemi.B2508545

基于多源扩散特征融合的行人目标检测算法研究*

张立成[1] 王 蕾[2] 宋逸菲[3] 谢耀华[4] 周竹萍[5] 陈春成[1]

(1. 长安大学信息工程学院 西安 710064;2. 空军工程大学空管领航学院 西安 710051;
3. 中航电测仪器(西安)有限公司 西安 710119;4. 招商局重庆交通科研设计院有限公司 重庆 400067;
5. 南京理工大学自动化学院 南京 210094)

摘 要:复杂环境下的行人检测是自动驾驶场景面临的重要技术挑战。提出一种基于红外与可见光多模态特征融合的行人目标检测方法。首先,针对跨模态特征融合难题,设计了一种基于扩散模型的红外与可见光图像融合算法,通过可逆全局捕获模块和轻量局部捕获模块实现多模态特征的有效提取与融合,提升融合图像的细节保真度和结构一致性。其次,针对低照度场景检测需求,提出了一种改进的 YOLOv8n 检测网络,通过引入 MobileNetV4 主干网络和颈部网络进行轻量化改进,并融入注意力机制和损失函数,增强关键特征提取能力和网络收敛性能。实验结果表明,所提方法在 LLVIP 和 MSRS 数据集上的融合图像质量指标显著优于现有方法,检测精度达到 95.3%,较基础模型提升 1.7%,模型参数量减少 42.6%,计算效率提升 17.4%。系统在复杂环境下的具备鲁棒性和实时性,为自动驾驶和智能安防应用提供了有效的技术解决方案。

关键词: 行人检测;图像融合;扩散模型;红外与可见光;YOLOv8

中图分类号: TP391.41;TN919.81 **文献标识码**: A **国家标准学科分类代码**: 510.50

Research on pedestrian target detection algorithm based on multi-source diffusion feature fusion

Zhang Licheng[1] Wang Lei[2] Song Yifei[3] Xie Yaohua[4] Zhou Zhuping[5] Chen Chuncheng[1]

(1. School of Information Engineering, Chang′an University, Xi′an 710064, China; 2. School of Air Traffic Control and Navigation, Air Force Engineering University, Xi′an 710051, China; 3. Zhonghang Electronic Measuring Instruments Co., Ltd., Xi′an 710119, China; 4. China Merchants Chongqing Communications Technology Research & Design Institute Co., Ltd., Chongqing 400067, China; 5. School of Automation, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: Pedestrian detection in complex environments presents a critical technical challenge for autonomous driving systems. This paper proposes a pedestrian target detection method based on multimodal feature fusion of infrared and visible light images. First, to address the challenge of cross-modal feature fusion, we propose an infrared-visible image fusion algorithm (DIVIF) based on a multi-source diffusion model. The algorithm incorporates a reversible feature decoupling module (IGCM) to capture target information distributions across modalities and extract large-scale structural features, along with a lightweight local capture module (LLCM) to enhance detail fidelity through attention-driven dynamic interaction. A global-local iterative optimization mechanism is introduced to achieve effective feature transfer and deep fusion. Second, for low-illumination scenarios, we propose an improved YOLOv8n detection network with lightweight architecture and attention mechanisms. The model integrates MobileNetV4 and GSConv into the backbone and neck networks for efficiency, while incorporating EMA attention to strengthen key feature extraction and MPDIoU loss to optimize convergence. Furthermore, a complete pedestrian detection system implementation is designed for practical applications. Experimental results demonstrate that the proposed method effectively fuses multimodal features, achieving precise detection in low-light and occluded

收稿日期:2025-07-08 Received Date: 2025-07-08

*基金项目:国家自然科学基金(52472314)、陕西省自然科学基金(2025JC-YBMS-408)、浙江省交通运输厅重大交通建设工程科研项目(2023-GCKY-09)资助

scenarios. Compared to mainstream detection networks, our algorithm outperforms in all evaluation metrics.

Keywords: pedestrian detection; image fusion; diffusion model; infrared and visible; YOLOv8

0 引 言

随着智能交通系统的快速发展,行人检测技术面临多维度挑战,在环境适应性方面,检测场景从理想光照扩展至夜间、低照度及恶劣天气等复杂环境,对系统的多模态感知能力提出更高要求;技术层面需要并行处理如红外与可见光等多源异构数据,对算法鲁棒性构成挑战;响应速度要求自动驾驶等场景具备毫秒级能力。传统单模态方法难以满足上述实际需求,亟需发展智能化的多模态融合技术[1]。为此,本文基于红外与可见光图像的互补特性,构建了一种鲁棒性更强的融合检测模型。红外图像通过热辐射信号突破光照限制,稳定捕捉目标轮廓;可见光图像则清晰呈现纹理与色彩细节。模型通过优化特征提取与多源信息整合机制,实现了多模态数据的协同感知增强。在本研究所提多源数据包括侧重于热辐射信息表征的红外模态数据,以及侧重于纹理细节信息表征的可见光模态数据。实验表明,所提方案显著提升了复杂场景下的行人检测精度。

行人检测技术经历了从传统手工特征到深度学习模型的演进。根据检测原理,现有方法可分为基于可见光、红外以及多模态融合 3 类方法,在环境适应性、精度检测和实时性等方面各具特点。

可见光行人检测技术经历了从传统方法到深度学习的演进[2]。传统方法中,Haar 特征结合 AdaBoost 分类器[3]、HOG 特征联合支持向量机(SVM)等方案计算高效,适用于早期简单场景,但难以应对复杂需求。深度学习方法中,基于区域的卷积神经网络(region-based convolutional neural,R-CNN)系列[4]精度突出,单次全局检测算法(you only look once,YOLO)系列凭借端到端范式更适配实时需求。其中轻量化的 DarkYOLOv8 融入 Transformer 注意力机制[5],适用于车载场景。针对低照度瓶颈,学界围绕 YOLOv8/YOLOv9 形成模块化改进体系,通过残差增强[6]、注意力与光照补偿融合[7]、双光自适应融合[8]等策略提升性能。但现有方法在极端低照度与复杂背景下仍存在特征退化与噪声抑制不足等问题。

红外行人检测技术凭借其全天候工作特性,在智能监控领域展现出独特优势[9]。传统方法如温度阈值分割[10]等在遮挡、杂波环境下性能大幅下降。基于卷积神经网络的方法逐步成为主流,Thermal YOLO 实现多模态协同处理[11],Transformer 编码器-解码器架构优化特征互补[12]。近年来,Transformer 与扩散模型进一步助力低信噪比与小目标识别,相关研究包括构建 U 型多尺度网络[13]、设计轻量化检测器[14]、Transformer 与 YOLO 融合方案[15]等。然而,当前技术在密集遮挡与极端天气下的适应性上仍有不足。

多模态融合技术为解决单模态行人检测瓶颈提供新路径,通过信息互补提升系统鲁棒性。近年研究围绕大模型形成“层次化跨模态对齐-轻量化适配融合”框架,采用大模型主体配合轻量化模块的低成本策略,兼顾特征表达能力与部署效率。具体通过可形变注意力轻量化网络优化红外弱目标检测[16]、多模态语义对齐与知识蒸馏提升跨场景适配性[17]、迁移学习强化低样本检测精度[18]、多传感器加权融合增强复杂环境鲁棒性[19]等方法,为多模态行人目标精准检测提供支撑。然而,现有技术在跨模态特征精细对齐、融合策略动态适配及极端环境下精度与效率的平衡等方面仍存在不足,制约了技术的实际应用。

1 系统总体设计思路

针对多光谱行人检测中特征融合与检测性能优化的核心问题,本文提出一种基于深度学习的改进方案。技术路线如图 1 所示。首先,在特征融合层面,引入扩散模型框架,设计可逆全局捕获模块处理跨模态特征对齐,配合轻量局部捕获模块进行细节特征提取,以应对多源图像融合的技术挑战;其次,在检测网络设计方面,以 YOLOv8n 为基础架构,引入轻量化模块与注意力机制,探索实时检测与精度提升的平衡路径;最后,通过系统集成将理论模型转化为工程应用,为智能安防与自动驾驶等实际场景提供技术支持。本研究从算法改进到系统实现,为复杂环境下的行人检测问题提供了一套完整的技术方案。

2 基于多源数据扩散模型的红外-可见光图像融合算法

针对单一模态图像的信息表达局限,本文提出一种基于扩散模型的融合算法,实现红外与可见光图像的高质量融合。算法通过构建包含可逆全局捕获模块和轻量局部捕获模块的反向融合网络,利用扩散过程的迭代去噪机制,有效保留红外图像的热辐射信息与可见光图像的纹理细节,在提升融合质量的同时降低模态差异干扰。

2.1 扩散模型框架

扩散模型是一种基于概率论的生成框架,其核心包

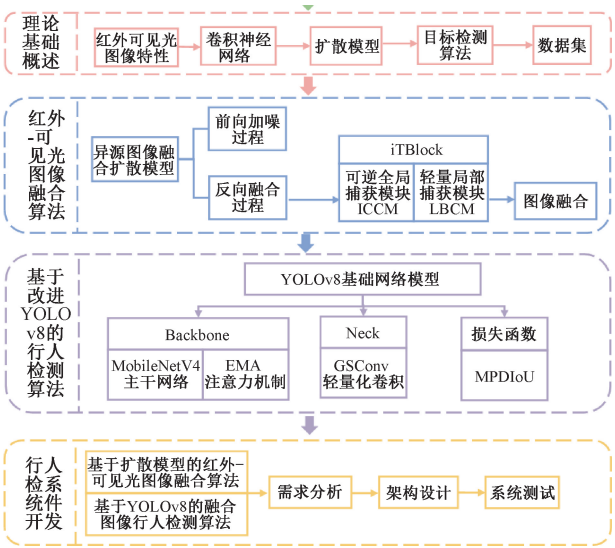

图 1　目标检测技术实现路径

Fig. 1　Architecture diagram for object detection implementation

含正向扩散和反向生成两个过程。模型最早由去噪扩散概率模型(denoising diffusion probabilistic model,DDPM)提出,通过模拟数据逐步退化为噪声,再逆向恢复生成新样本,在图像生成、文本生成等领域表现优异。

1)正向扩散过程

正向扩散过程从原始图像开始,逐步添加方差为 $\beta_t(t=1,2,\cdots,T)$ 的高斯噪声,使得图像逐渐退化到接近纯高斯噪声,得到服从标准正态分布。该扩散过程构成一个马尔科夫链,未来状态只因当前状态变化而变化,与历史状态无关。每步的扩散过程表示为:

$$q(x_t\mid x_{t-1})=\mathrm{N}\left(x_t;\sqrt{a_t}x_{t-1},(1-\alpha_t)\mathrm{I}\right) \tag{1}$$

$$q(x_{0:T})=q(x_0)\prod_{t=1}^{T}q(x_t\mid x_{t-1}) \tag{2}$$

式中:$\alpha_t=1-\beta_t$,β_t 通常满足 $0<\beta_1<\beta_2<\cdots<\beta_T<1$;$N(\cdot)$ 为标准正态分布。通过计算得到任意时刻的加噪后图像,为反向过程做准备。

2)反向生成过程

反向生成过程即去噪过程,是扩散模型的核心,从初始噪声图像(服从 $N(0,1)$ 的标准高斯分布)出发,多次去噪迭代还原出原始数据。假设逆向扩散过程用马尔科夫链式表示,反向过程表示为:

$$p_\theta(x_{t-1}\mid x_t)=N(x_{t-1};\mu_\theta(x_t,t),\sum\nolimits_\theta(x_t,t)) \tag{3}$$

$$p(x_{0:T})=p(x_T)\prod_{t=1}^{T}p(x_{t-1}\mid x_t) \tag{4}$$

式中:θ 为神经网络中通过训练学习的参数;$\mu_\theta(x_t,t)$ 和 $\sum_\theta(x_t,t)$ 分别为第 t 步分布的均值和方差;逆向扩散过程中的转移分布 $p(x_{t-1}\mid x_t)$ 近似为前向扩散过程的转移分布 $q(x_{t-1}\mid x_t,x_0)$。

假设在给定 x_t 的情况下,并且得到 x_{t-1} 的分布,即能够推导出 $q(x_{t-1}\mid x_t)$,从随机的噪声图片出发,通过逐步采样生成类似原始图像的去噪图像。因此,$q(x_{t-1}\mid x_t,x_0)$ 表示为:

$$q(x_{t-1}\mid x_t,x_0)=q(x_t\mid x_{t-1},x_0)\frac{q(x_{t-1}\mid x_0)}{q(x_t\mid x_0)} \tag{5}$$

式中:$q(x_{t-1}\mid x_0)$、$q(x_t\mid x_0)$、$q(x_t\mid x_{t-1},x_0)$ 分别服从以下正态分布,具体表示为:

$$q(x_{t-1}\mid x_0)=\sqrt{\bar{a}_{t-1}}x_0+\sqrt{1-\bar{a}_{t-1}}z\sim N\left(\sqrt{\bar{a}_{t-1}}x_0,(1-\bar{a}_{t-1})I\right) \tag{6}$$

$$q(x_t\mid x_0)=\sqrt{\bar{a}_t}x_0+\sqrt{1-\bar{a}_t}z\sim N\left(\sqrt{\bar{a}_t}x_0,(1-\bar{a}_t)I\right) \tag{7}$$

$$q(x_t\mid x_{t-1},x_0)=\sqrt{a_t}x_{t-1}+\sqrt{1-a_t}z\sim N\left(\sqrt{a_t}x_{t-1},(1-a_t)I\right) \tag{8}$$

其中,z 服从标准正态分布,将以上 3 个分布代入式(5)中,展开后如式(9)所示。

$$q(x_t\mid x_{t-1},x_0)\propto\exp\left(-\frac{1}{2}\left(\frac{\left(x_t-\sqrt{a_t}x_{t-1}\right)^2}{\beta_t}+\frac{\left(x_{t-1}-\sqrt{\bar{a}_{t-1}}x_0\right)^2}{1-\bar{a}_{t-1}}-\frac{\left(x_t-\sqrt{\bar{a}_t}x_0\right)^2}{1-\bar{a}_t}\right)\right)=$$

$$\exp\left(-\frac{1}{2}\left(\left(\frac{a_t}{\beta_t}+\frac{1}{1-\bar{a}_t}\right)x_{t-1}^2-\left(\frac{2\sqrt{a_t}}{\beta_t}x_t+\frac{2\sqrt{\bar{a}_{t-1}}}{1-\bar{a}_{t-1}}x_0\right)x_{t-1}+C(x_t,x_0)\right)\right) \tag{9}$$

式中:C 表示仅与 x_0 和 x_t 相关的常数项。

对式(9)进一步推导,高斯分布表达式为:

$$\mathrm{N}(x;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{10}$$

均值 $\widetilde{\mu}_t(x_t,x_0)$、方差 σ_t^2 表达式为:

$$\widetilde{\mu}_t(x_t,x_0)=\frac{\sqrt{a_t}(1-\bar{a}_{t-1})}{1-\bar{a}_t}x_t+\frac{\sqrt{\bar{a}_{t-1}}\beta_t}{1-\bar{a}_t}x_0 \tag{11}$$

$$\sigma_t^2=\frac{(1-a_t)(1-\bar{a}_{t-1})}{1-\bar{a}_t}I \tag{12}$$

方差 σ_t^2 通过参数计算得到常数。均值 $\widetilde{\mu}_t$ 的计算中只有 x_0 为未知数,通过神经网络训练噪声估计 x_0。

当神经网络模型训练完成并能够准确预测噪声时,就可以通过逐步采样,将随机噪声图像还原为原始图像。其中采样公式可以表示为:

$$x_{t-1}=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t,t)\right)+\sigma_t z \tag{13}$$

式中:z 为随机噪声;$x_T\sim N(0,1)$。

扩散模型为红外-可见光图像融合提供了系统性解

决方案:其渐进融合机制平衡了红外目标显著性与可见光纹理,避免信息失衡;基于概率生成框架的自适应权重调节增强了场景适应性;通过可逆变换与信息引导模块的结合,在保留关键信息的同时抑制冗余,实现语义与结构的一致性。该框架为跨模态融合提供了兼具理论严谨性与实践可行性的新途径。

2.2 多源扩散模型图像融合算法实现

本文提出一种基于扩散模型的红外-可见光图像融合算法(diffusion-based infrared-visible image fusion model, DIVIF)。该方法采用多阶段交替优化策略,通过增强红外与可见光图像的细节与特征,实现高质量的融合图像生成。DIVIF 由前向加噪扩散和反向连续融合两个核心过程:前向过程逐步向输入图像添加高斯噪声,直至其转化为纯噪声;反向过程则在输入图像的引导下,逐时间步执行多源图像融合,逐步重构出融合结果。

1)前向加噪

对于给定可见光和红外图像对,DIVIF 模型的目标是通过逐步提取和融合多模态特征,生成高质量融合图像 X。其正向扩散过程通过逐步添加高斯噪声,将目标融合图像分布逐渐退化至纯高斯分布,过程表示为:

$$q(X_t \mid X_{t-1}) = N\left(X_t; \sqrt{1-\beta_t} X_{t-1}, \beta_t \boldsymbol{I}\right) \tag{14}$$

式中:$t \in [0,T]$;X_t 表示第 t 步的噪声图像;X_{t-1} 表示第 $t-1$ 步的图像;β_t 是噪声衰减系数;$\boldsymbol{I}$ 是单位矩阵;方差 $1-\beta_t$ 控制噪声的添加强度,随着时间的增加,噪声逐步主导,使 X_t 趋近于高斯分布。

前向加噪的设计具有重要意义,首先,将复杂图像分布转化为简单高斯分布,为反向融合提供明确目标;其次,通过可逆噪声添加为反向生成提供理论支撑;平衡特征保留与分布退化的关系,避免关键信息过早丢失。

在 DIVIF 模型中,噪声衰减系数 β_t 的设计遵循模态特征保护与分布协同原则。红外图像热辐射分布平缓,对强噪声具备较强抗干扰能力。可见光图像纹理细节特征,过度噪声易导致关键细节丢失。基于此,β_t 的初始范围设定为[0.000 1,0.02],并满足 $\beta_1 < \beta_2 < \cdots < \beta_T$($T=1\,000$ 为扩散总步数)。前 300 步的 β_t 增长速率放缓,β_t 从 0.000 1 逐步增至 0.005,确保红外图像的热辐射峰值和可见光图像的关键纹理在噪声添加初期不被破坏;后续 β_t 加速增长至 0.02,快速推动双模态数据向统一高斯分布收敛,避免了融合过程中模态差异导致的分布偏移。为进一步提升多模态特征的兼容性,通过引入模态特征重要性因子 ω_t 动态优化:

$$\omega_t = \frac{\lambda_{IR} \times S_{IR}(X_t - 1) + \lambda_{VIS} \times S_{VIS}(X_t - 1)}{\lambda_{IR} \times \lambda_{VIS}} \tag{15}$$

$$\beta'_t = \beta_t \times \omega_t \tag{16}$$

式中:$\lambda_{IR}=0.4$、$\lambda_{VIS}=0.6$ 为模态权重系数;$S_{IR}(X_t-1)$ 表示红外图像在 $t-1$ 步的特征显著性,通过计算热辐射区域的灰度方差量化;$S_{VIS}(X_t-1)$ 表示可见光图像在 $t-1$ 步的特征显著性,通过 Sobel 梯度算子计算纹理细节丰富度。$S_{IR}(X_t-1)$ 当较高,例如红外目标轮廓清晰时,ω_t 偏向抑制 β 增长,减少噪声对热辐射特征的干扰。$S_{VIS}(X_t-1)$ 较高,如可见光纹理密集时,ω_t 适度提升 β_t,但仍控制在 0.02 以内,避免细节过度模糊。β_t 的自适应调整与扩散过程的模态融合状态深度绑定,通过实时监测双模态特征的分布一致性指标。通过上述取值依据、优化策略与自适应调整逻辑,β_t 能够精准匹配红外-可见光的模态特性,在保证数据分布顺利退化的同时,最大化保留双模态的核心特征,为后续反向融合阶段的特征互补奠定基础。

2)反向融合

本文提出的 DIVIF 方法基于反向扩散过程,其核心单元(integrated base and local block, IBL-Block)由可逆全局捕获模块(invertible global capture module, IGCM)和轻量局部捕获模块(lightweight local capture module, LLCM)构成,实现多模态特征融合。IGCM 提取并增强双模态图像的全局结构特征,LLCM 则通过注意力机制动态交互局部细节特征(如边缘与纹理)。两模块在反向去噪中交替优化,逐步细化特征并实现跨模态互补,最终生成兼顾热辐射信息与纹理细节的高质量融合图像。

(1)可逆全局捕获模块 IGCM

可逆神经网络(invertible neural network, INN)针对红外与可见光图像融合任务中,细节特征(如边缘、纹理等)保留的关键需求,本文设计基于可逆网络的 IGCM 模块,通过双向映射架构实现多模态特征的无损传递与融合,其结构如图 2 所示。

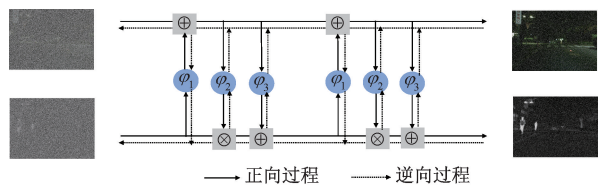

图 2 IGCM 模块结构

Fig. 2 Structure diagram of IGCM module

在扩散时间步 t,IGCM 模块接收来自前一时间步的红外图像特征 X_t^{ir} 和可见光图像特征 X_t^{vi} 作为输入,通过交互提取并融合局部特征。具体流程如下:在正向过程中,通过非线性映射函数编码,通过 N 个可逆块。在第 i 个可逆块时,输入为 $X_t^{ir(i)}$ 和 $X_t^{vi(i)}$,输出为 $Y_t^{ir(i+1)}$ 和 $Y_t^{vi(i+1)}$,具体表示为:

$$Y_t^{ir(i+1)} = X_t^{ir(i)} + \varphi_1(X_t^{vi(i)}) \tag{17}$$

$$Y_t^{vi(i+1)} = X_t^{vi(i)} \odot (\varphi_2(Y_t^{ir(i+1)})) + \varphi_3(Y_t^{ir(i+1)}) \tag{18}$$

式中:φ_1、φ_2 和 φ_3 是可学习的神经网络,用于提取与融

合特征。对于 φ_1、φ_2 和 φ_3，采用 MobileNetV 中的瓶颈残差块(bottleneck residual block，BRB)作为网络架构。在 N 个可逆块之后，得到 X_{t-1} 和信息损失 L，IGCM 能够有效提取模态间互补信息。在逆向过程中，同样由 N 个可逆块组成，并与正向过程共享相同的参数，但信息流的方向与正向相反。通过这一过程，模型能够将正向过程的输出重新映射回输入，确保融合期间的信息损失最小化。

(2)轻量局部捕获模块 LLCM

本研究基于 Transformer 架构，该模型通过自注意力机制实现了对序列数据的高效建模。本文采用轻量化改进的 Lite Transformer 架构[20]，在保留编码器-解码器基本框架的基础上，通过分层编码器实现多尺度特征提取，并采用无位置编码设计避免分辨率依赖问题。模型引入卷积自注意力(convolutional self-attention，CSA)机制增强底层局部特征表征能力，结合递归膨胀自注意力(RASA)机制实现多尺度高层语义特征学习。这两种注意力机制的协同工作，使模型在保持轻量化特点的同时，具备更强的多尺度特征学习能力。模型如图 3 所示。

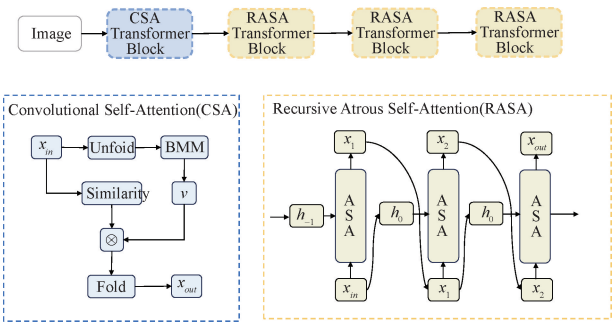

图 3　Lite Transformer 框架图

Fig. 3　Framework diagram of lite transformer

CSA 是一种融合卷积神经网络和自注意力机制的特征提取方法，通过卷积提取局部特征，自注意力建模全局关系，能够同时处理局部和全局信息，可以有效提取图像特征。本文提出了一种新型的卷积自注意力方法，不同于传统的 CNN 与自注意力直接结合，通过基于窗口的方式将自注意力嵌入到 CNN 中，以增强长距离依赖建模能力，同时保留卷积的局部特征提取优势。CSA 的操作图如图 4 所示。

设输入和输出特征向量分别为 x 和 y，其中 d 表示通道数。定义 i 为空间位置索引，卷积操作通过滑动窗口进行计算。在每个窗口中，卷积公式表示为：

$$y_i = \sum_{j \in N(i)} a_{ij} \boldsymbol{W}_{ij} x_j \tag{19}$$

式中：$\boldsymbol{W}_{ij}$ 为投影矩阵；$N(i)$ 表示以位置 i 为中心的局部邻域，且 $N(i) = k \times k$；k 为卷积核大小。

在自注意力计算中，需要 3 个投影矩阵 $\boldsymbol{W}_q$、$\boldsymbol{W}_k$、$\boldsymbol{W}_v$ 来计算 query、key 和 value，是 $d \times d$ 维度的矩阵。自注意

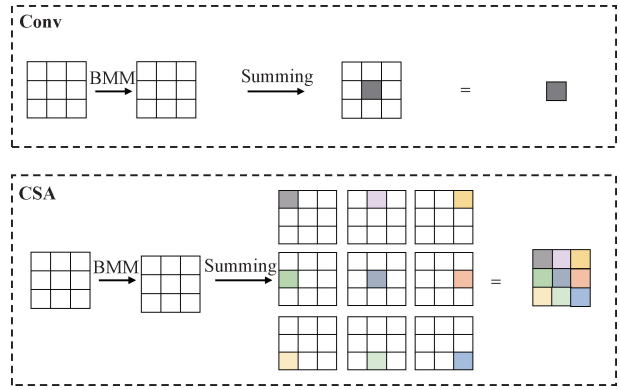

图 4　Conv 和 CSA 对比

Fig. 4　Comparative diagram of Conv and CSA

力计算公式表示为：

$$y_i = \sum_{j \in N(i)} a_{ij} \boldsymbol{W}_v x_j \tag{20}$$

$$a_{ij} = \frac{e^{(\boldsymbol{W}_q x_i)^{\mathrm{T}} \boldsymbol{W}_k x_j}}{\sum_{z \in N(i)} e^{(\boldsymbol{W}_q x_i)^{\mathrm{T}} \boldsymbol{W}_k x_z}} \tag{21}$$

式中：a_{ij} 是一个缩放因子，用来控制不同位置下的输入 x 的比例，这个因子通过归一化，使得因子之和为 1。将自注意力和卷积推广为统一的卷积自注意力，其公式表示为：

$$y_i = \sum_{j \in N(i)} a_{ij} \boldsymbol{W}_{ij} x_j \tag{22}$$

本文提出 RASA 模块，结合递归机制与 ASA 以增强表征能力。ASA 部分通过空洞卷积扩展，同时采用权重共享和组卷积策略来降低参数量。不同尺度的特征通过加权求和融合，权重由自校准机制根据激活强度动态确定，并通过 SiLU 函数实现非线性变换。这种设计使 self-attention 中任意两个空间位置之间的 query 和 key 的相似度计算能够融入多尺度信息，从而提升建模能力，公式表示为：

$$Q = \sum_{r \in \{1,3,5\}} SiLU(Conv(\hat{Q}, \boldsymbol{W}_q^{k=3}, r, g = d)) \tag{23}$$

$$\hat{Q} = Conv(X, \boldsymbol{W}_q^{k=1}, r = 1, g = 1) \tag{24}$$

式中：d 为特征通道数；k、r、g 分别表示卷积核的大小，膨胀速度和群数。

对于轻量级模型，在不增加参数量的情况下增加模型深度，引入递归方法。与传统的卷积神经网络递归方法不同，本文将递归机制应用于自注意力中，提出了递归自注意力，其设计遵循标准循环网络的结构，公式表示为：

$$x_{t+1} = ASA(F(X_t, h_{t-1})) \tag{25}$$

$$h_{t-1} = X_{t-1} \tag{26}$$

$$X_t = ASA(F(X_{t-1}, h_{t-2})) \tag{27}$$

式中：h_t 表示第 t 步的隐藏状态；X 为输入特征。

3)网络优化目标

DIVIF 的目标是通过端到端的学习过程,经过一系列的融合步骤来生成融合图像。因此,在损失函数设计方面首先考虑了传统的内容损失,包括强度和纹理损失。强度损失用于约束融合结果的全局亮度一致性,以确保重建图像在整体强度上的准确性;纹理损失则用于保留和强化图像中的局部纹理细节,确保融合图像在纹理层面的细节丰富度。内容损失计算公式表示为:

$$L_{content} = L_{int} + \alpha L_{texture} \tag{28}$$

式中:α 为已知参数,用于维持两个损失函数之间的平衡。

强度损失 L_{int} 和纹理损失 $L_{texture}$ 可表示为:

$$L_{int} = \frac{1}{HW} \| I_f - \max(I_{ir}, I_{vi}) \|_1 \tag{29}$$

$$L_{texture} = \frac{1}{HW} \| \nabla I_f | - \max(| \nabla I_{ir} |, | \nabla I_{vi} |) | \|_1 \tag{30}$$

式中:H 表示图像的高度;W 为图像宽度;$\| \cdot \|_1$ 为 L_1 范数;∇ 表示 Sobel 梯度算子,通过计算图像空间梯度来量化局部纹理特征。

3 基于改进 YOLOv8 的融合图像行人检测算法

本研究基于 YOLOv8 算法,通过骨干网络和颈部网络的轻量化改进,结合注意力机制和损失函数优化,提升多模态融合图像的行人检测实时性和准确性。重点增强多模态特征融合与行人定位能力,在保证检测精度的同时降低计算复杂度,以适应嵌入式设备的性能需求。

3.1 目标检测算法

两阶段目标检测算法采用分阶段处理策略,以 Faster R-CNN 为代表,通过特征提取、区域建议和检测子网络实现高精度检测,但是计算复杂度高。而单阶段方法,通过端到端设计整合定位与识别,显著提升效率,更适用于实时场景。YOLOv8 版本由 4 个功能层构成:Input 层负责图像预处理,Backbone 基于改进的 CSPDarkNet 架构,集成 Conv、C2f 及 SPPF 模块以完成特征提取与多尺度融合;Neck 层沿用 FPN+PAN 结构进一步融合多级特征;Head 层采用解耦式 Anchor-Free 检测头。YOLOv8 通过上述优化,在检测精度、计算效率与小目标识别之间取得了良好平衡,成为当前目标检测领域的重要方案。

3.2 YOLOv8 融合图像行人检测改进算法

尽管 YOLOv8 在目标检测中性能优异,但模型参数量大计算效率受限,难以满足移动端实时检测需求。为此,本研究聚焦于主干网络和颈部网络的轻量化设计,通过优化模型结构复杂度并引入注意力机制,在显著提升计算效率的同时,保持较高的检测精度。

1)主干网络轻量化设计

本研究采用 MobileNetV4 作为 YOLOv8n 的主干网络,旨在实现模型轻量化、提升计算效率的同时保持检测精度。MobileNetV4 引入了多头查询注意力(multi query attention,MQA)模块,该机制通过并行处理多个查询向量,实现对多维度特征的差异化关注。具体而言,多个查询头分别与共享的键值对交互,生成相应的注意力权重,最终融合为多路特征表示。此设计兼顾了特征表达能力与计算效率,尤其适用于多任务多目标场景。

MQA 模块中,每个查询头 $\boldsymbol{Q}_j$ 的计算公式为:

$$attention_j = softmax(X\boldsymbol{W}_{Q_j})(SR(X)\boldsymbol{W}_K)^{\mathrm{T}} \frac{1}{\sqrt{d_k}} (SR(X)\boldsymbol{W}_V) \tag{31}$$

式中:X 是输入特征图;$\boldsymbol{W}_{Q_j}$ 是查询头的权重矩阵;$\boldsymbol{W}_k$ 和 W_v 分别是键和值的权重矩阵;d_k 是键的维度;SR 表示空间降维操作,用于降低键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$ 的分辨率。

MQA 模块通过拼接来自不同查询头的注意力结果计算最终输出,计算公式为:

$$Mobile\_MQA(X) = Concat(attention_1, \cdots, attention_n)\boldsymbol{W}^O \tag{32}$$

式中:$attention_n$ 表示每个查询头的注意力结果,拼接后的结果通过一个线性变换矩阵 $\boldsymbol{W}^O$ 进行变换,得到最终输出结果。

MobileNetV4 根据计算需求提供 Small、Medium 和 Large 3 种规模变体。为满足轻量化目标,选用计算开销最小的 MobileNetV4-Small,其在行人检测任务中具备以下适配优势:行人轮廓、肢体结构等中低层语义特征可通过深度可分离卷积有效提取,避免深层复杂网络带来的特征冗余;计算量约为 Large 版本的 1/3,可在车载终端等资源受限场景实现低延迟实时监测;MQA 模块能够强化头部、躯干等关键区域的特征权重,提升复杂背景下的行人区分能力。

网络设计中引入的 ExtraDW 结构进一步增强模型对低分辨率及多尺度行人的特征提取鲁棒性,在控制计算复杂度的同时保障了检测精度。改进后的网络结构如图 5 所示。

2)颈部网络轻量化设计

在计算机视觉任务中,图像特征提取过程中伴随着高维通道编码与语义信息损失问题。GSConv 通过保持通道间连接缓解语义损失,但需权衡其引入的计算开销。在特征融合阶段,特别是特征图通道数最大且空间分辨率较低时,GSConv 能有效减少冗余信息,降低计算负担。GSConv 采用多分支特征融合架构:首先通过标准卷积提

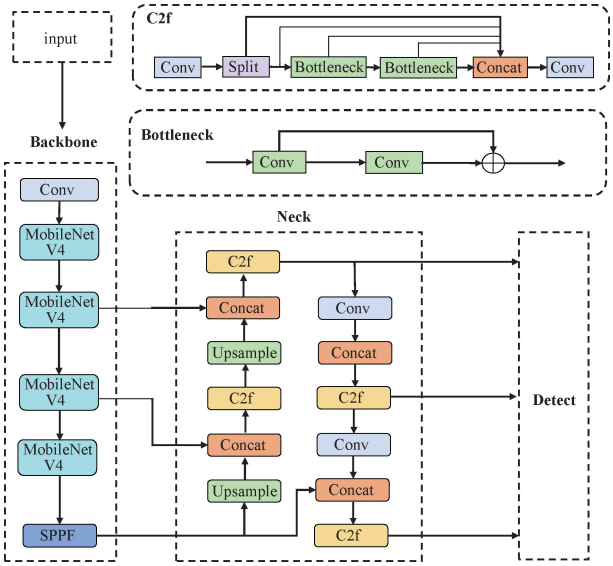

图 5　主干网络引入 MobileNetV4 的 YOLOv8n 框架

Fig. 5　YOLOv8n framework with MobileNetV4 backbone

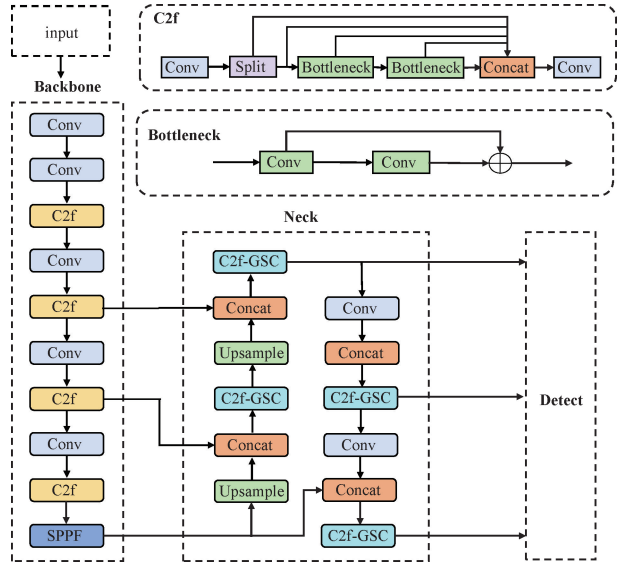

图 6　颈部网络引入 GSConv 的 YOLOv8n 框架

Fig. 6　YOLOv8n framework with GSConv-Enhanced neck network

取全局基础特征,再结合深度可分离卷积进行轻量化局部特征提取,最终实现跨通道高效交互。该设计平衡了全局表征和局部特征提取能力,在保证高效率的同时增强语义融合效果。GSConv 的计算量表示为:

$$FLOPs_{GSConv} = H \times W \times C_1 \times C_2 \times \frac{K \times K}{G} \tag{33}$$

式中:W、H 分别为输出特征图的宽度和高度;C_1、C_2 分别是输入特征图和输出特征图的通道数;K 为卷积核大小;G 表示分组数。

本文将 GSConv 与 C2f 模块结合,提出 C2f-GSC 轻量化模块并应用于 YOLOv8n 颈部网络。该模块针对行人检测任务进行优化,其多分支结构可同步提取全局姿态与局部细节,提升对姿态多变与遮挡目标的识别能力。GSConv 保持通道连接的特性有助于维护行人轮廓的语义一致性,减少轮廓断裂现象,适应复杂背景下的精确定位需求。模块的轻量化设计与 MobileNetV4-Small 主干网络协同,通过 C2f 残差连接强化特征传播,GSConv 跨通道交互降低冗余计算,实现在密集行人场景中高效且准确的检测,有效平衡精度与效率。颈部网络引入 C2f-GSC 的 YOLOv8n 框架如图 6 所示。

3)引入注意力机制的轻量化 YOLOv8 网络

注意力机制通过动态权重分配输入数据,分为空间域、通道域、时间域和混合域等类型。在目标检测中,该机制优化特征选择,尤其适用于红外与可见光融合的行人检测,能过滤噪声并提升检测性能。

(1)SE(squeeze-and-excitation)注意力机制

SE 注意力机制通过特征通道重标定增强网络表征能力。其压缩阶段采用全局平均池化,将 $H \times W \times C$ 的特征图 U 压缩成 $1 \times 1 \times C$ 的向量 $\boldsymbol{F}_{sq}$,通过聚合全局特征构建全局表征,量化各通道在图像中的重要性,计算表示为:

$$Z_c = \boldsymbol{F}_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{34}$$

式中:H 和 W 为特征图宽度和高度;$u_c(i,j)$ 为特征图的一个像素点。

SE 注意力机制的激励机制采用瓶颈式全连接结构建模通道间依赖关系,经升维全连接层恢复通道维度,通过 Sigmoid 函数生成归一化权重,动态加权增强关键特征表示为:

$$s = F_{ex}(z, W) = \sigma(g(z, W)) = \sigma(\boldsymbol{W}_2 ReLU(\boldsymbol{W}_1 z)) \tag{35}$$

式中:$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 分别表示全连接层的可学习参数矩阵,Z 表示经过压缩操作后的结果;σ 是 Sigmoid 激活函数,输出值的范围在[0,1]之间。

SE 模块将激励产生的通道权重与输入特征逐通道相乘,输出校准后的特征 $\tilde{X}_c$,增强关键信息并抑制冗余,显著提升模型表达能力,计算表示为:

$$\tilde{X}_c = F_{scale}(u_c, u_s) = s_c \times u_c \tag{36}$$

式中:s_c 表示各通道的权重数;u_c 表示未经压缩的特征图。

(2)CBAM 注意力机制 P 卷积块注意力模块(convolutional block attention module,CBAM)由通道注意力模块(channel attention module,CAM)和空间注意力模块(spatial attention module,SAM)组成。CAM 通过聚合全局信息来强化任务相关的特征通道响应,SAM 精确

定位目标的关键区域。这种双分支设计使网络自适应聚焦关键信息,有效抑制干扰,显著提升特征提取效率,且不增加计算复杂度,如图 7 所示。

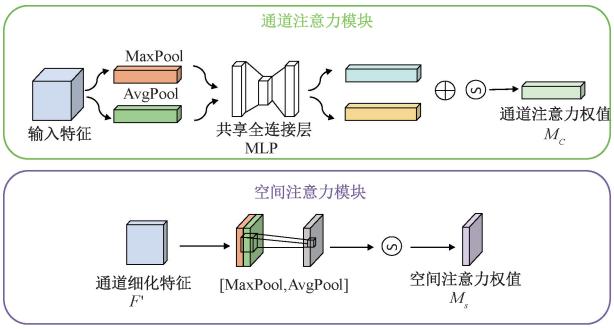

图 7 CBAM 注意力模块详细结构

Fig. 7 Detailed architecture diagram of CBAM

通道注意力 M_c 机制通过多阶段处理实现特征通道的加权:首先对输入特征图 F 执行全局池化,提取显著特征;通过 Sigmoid 函数生成归一化的通道权重系数 c_M;通过 Hadamard 乘积实现特征通道的自适应校准。该机制在保持空间结构的同时增强特征选择性,M_c 表示为:

$$M_c = \sigma(MLP(avgpool(F)) + MLP(maxpool(F))) = \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \quad (37)$$

空间注意力 M_s 机制通过多级特征处理实现加权:首先经自适应卷积生成空间权重图,将 Sigmoid 函数归一化后与输入特征相乘,实现空间维度的特征增强,突出关键区域并抑制背景噪声,M_s 表示为:

$$M_s(F) = \sigma(f^{7\times7}([avgpool(F);maxpool(F)])) = \sigma(f^{7\times7}([F^s_{avg};F^s_{max}])) \quad (38)$$

(3)高效多尺度注意力(EMA)机制

EMA 注意力机制引入历史信息实现高效特征建模,在保留通道关键特征的同时优化计算效率。EMA 显著增强模型稳定性和长期依赖捕捉能力,如图 8 所示。

对于输入特征图,H 和 W 分别为特征图的高和宽;C 为输入通道数。首先沿通道维度划分为 G 个互不重叠的子组。随后,通过 3 个并行卷积分支提取特征:两个 1×1 卷积和一个 3×3 卷积。各分支输出分别经过横向(X Avg Pool)与纵向(Y Avg Pool)的全局平均池化处理,再进行特征聚合,生成注意力加权特征图。该多尺度融合策略增强了模型对全局上下文的建模能力。

综上所述,本文以 YOLOv8n 为基础,在优化主干网络和颈部网络的基础上,为缓解轻量化带来的精度损失,在主干网络末端集成了 SE、CBAM 和 EMA 模块。改进后的网络结构如图 9 所示。

4)损失函数设计

YOLOv8 的损失函数设计上延续 YOLOv5 的完全交

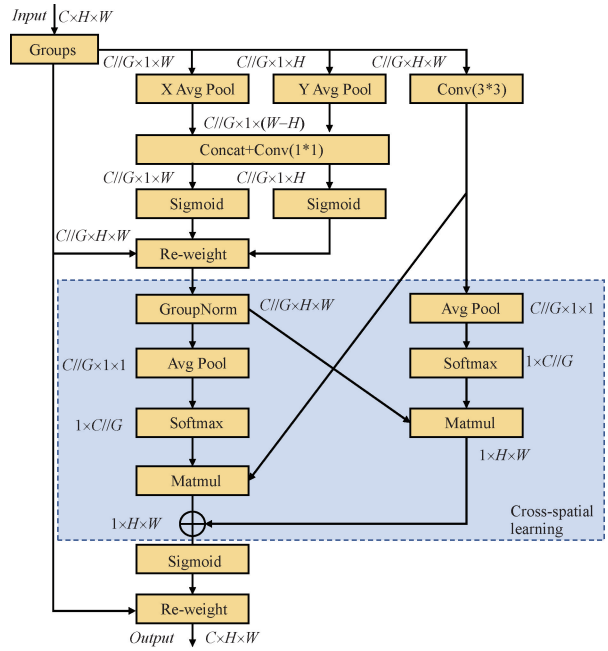

图 8 EMA 注意力模块

Fig. 8 EMA attention mechanism diagram

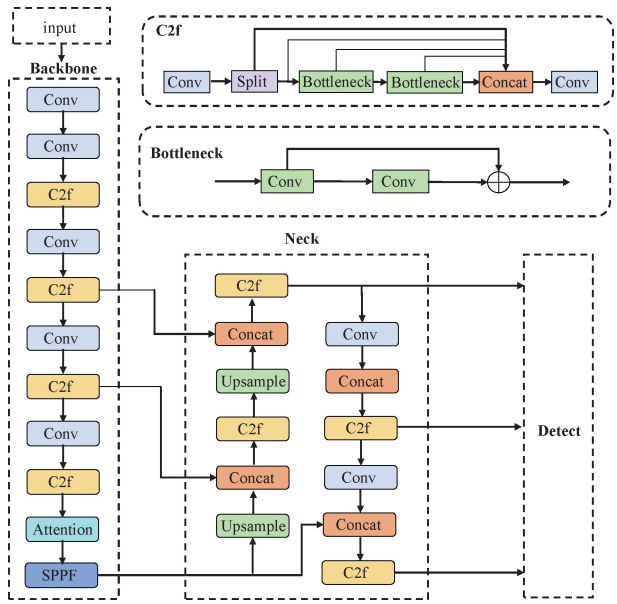

图 9 主干网络引入注意力机制的 YOLOv8n 框架

Fig. 9 YOLOv8n framework with attention-enhanced backbone network

并比损失(complete intersection over union loss,CIoU)与无锚框预测策略,并引入分布焦点损失(distribution focal loss,DFL)机制提升目标中心定位能力。但在遮挡和边界模糊场景下,CIoU 对几何形变目标适应性不足,DFL 对复杂定位的精度有限。为此,本文提出最小点距离交并比损失(minimum point distance intersection over union,

MPDIoU)边界框回归损失,该损失综合考虑重叠区域、中心点距离以及对角线最小点距离,以提升复杂场景下的回归精度。

假设输入图像的宽度与高度为 w 和 h;真实框记为 B_{gt},坐标为 $(x_c^{gt},y_c^{gt},w^{gt},h^{gt})$;预测框表示为 B_{prd},预测框坐标为 (x_c^p,y_c^p,w^p,h^p)。选取两框对角线顶点对,其最小欧氏距离可表示为:

$$d_{\min}=\min\left(\sqrt{(x_1^{gt}-x_2^p)^2+(y_1^{gt}-y_2^p)^2},\sqrt{(x_2^{gt}-x_1^p)^2+(y_2^{gt}-y_1^p)^2}\right) \tag{39}$$

式中:(x_1,y_1) 和 (x_2,y_2) 分别为框的左上角与右下角坐标。

损失函数的计算式为:

$$L_{MPDIoU}=1-MPDIoU=1-\left[\frac{|B^{gt}\cap B^{pred}|}{B^{gt}\cup B^{pred}}-\lambda\times\frac{d_{\min}}{d_{\max}}\right] \tag{40}$$

式中:$\lambda\in[0,1]$。

4　实验与仿真

4.1　改进式融合算法仿真实验

1)实验环境设置

为了证实本方法的有效性,改进式融合算法仿真实验将在 LLVIP[21] 和 MSRS[22] 数据集上展开。本研究的算法实现基于 Python 编程语言,采用 PyTorch 深度学习框架进行神经网络的构建与训练。开发环境选用 PyCharm 作为集成开发工具,在搭载 NVIDIA GeForce RTX 3060 GPU 的系统上完成。具体参数设置如表 1 所示。

表 1　网络训练参数设置

Table 1　Neural network training configuration

配置名称	配置信息
初始学习率	1×10^{-3}
优化方法	Adam
批量大小	2
步长	8
迭代次数	100
衰减率	2×10^{-4}

2)红外-可见光数据集本文选用 LLVIP 和 MSRS 多模态数据集进行实验。LLVIP 包含 16 836 对严格配准的可见光与红外图像,主要覆盖低光/夜间场景,具有超高分辨率和时空对齐特性。从中选取 3 400 对昼夜场景图像用于实验。MSRS 数据集包含 1 444 对配准图像(昼夜各半),采用暗通道增强技术提升红外对比度,并包含精细的道路目标标注。两个数据集的多场景特性为跨模态检测研究提供了全面验证基础[23]。

3)图像融合评价指标

本文采用 6 种客观指标综合评价融合图像质量。从信息量、纹理结构、相关性和视觉效果等多角度克服单一指标的局限性。标准差(standard deviation,SD)反映数据离散程度,可衡量图像对比度与信息丰富度;视觉保真度(visual information fidelity,VIF)量化原始内容的保留程度,评估感知信息的损失;平均梯度(average gradient,AG)衡量细节丰富度和纹理清晰度,其值与图像质量正相关;信息熵(information entropy,EN)评估图像信息量;基于梯度的融合性能(gradient-based fusion performance,Qabf)聚焦源图像的显著性特征在融合结果中的保留程度;空间频率(spatial frequency,SF)用于表征图像在空间域中的信息活跃程度。该多维度指标体系为算法优化提供了全面、可靠的依据。

4)消融实验分析

针对改进的扩散模型中反向扩散网络中的 IGCM、LLCM 模块在 MSRS 数据集上开展实验。构建 3 个变体模型 VU1、VU2、VU3,使用常规的 U-Net 结构分别替换 IGCM、LLCM 模块,结果如表 2 所示。

实验结果表明,DIVIF 在所有指标上均显著优于对比模型,特别是在细节保留和结构恢复方面表现突出。这得益于其可逆结构设计有效减少了信息损失,以及特征提取模块对多模态信息的有效捕捉。定量分析证实,本文方法在各项指标上均优于现有方法,显著提升了图像融合质量。

为验证可逆全局和局部捕获模块的性能优势,本研究对比了不同模块变体与 DIVIF 的可视化结果。如图 10 所示,红色框突出细节特征,绿色框标记行人区域。实验表明,DIVIF 融合图像既保留了行人特征,又融合了可见光细节,验证了其有效性和实用性。

表 2　模块消融实验定量对比

Table 2　Quantitative ablation study on module components

算法模型	算法模块	SD	VIF	AG	EN	Qabf	SF
VU1	U-Net、LLCM	43.846 1	0.860 3	3.648 3	5.513 5	0.483 6	7.460 2
VU2	IGCM、U-Net	39.690 0	0.867 5	3.812 6	6.145 8	0.525 3	8.498 9
VU3	U-Net、U-Net	34.865 2	0.846 9	3.548 2	5.054 6	0.417 3	6.164 2
DIVIF	IGCM、LLCM	**63.247 7**	**0.908 5**	**4.088 0**	**7.128 2**	**0.682 0**	**12.509 3**

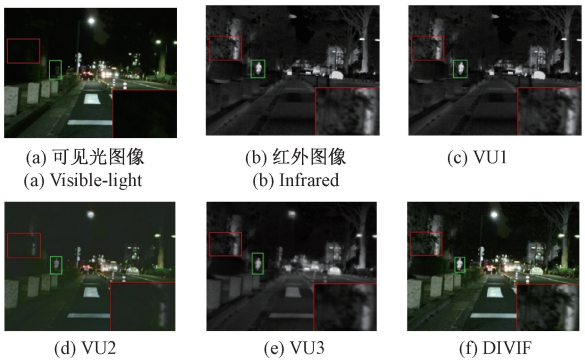

图 10 消融实验对比图

Fig. 10 Ablation study comparison diagram

5)对比实验分析

(1)定性分析

在 MSRS 和 LLVIP 数据集上对 DenseFuse[24]、FusionGAN[25]、RFN-Nest[26]、SwinFusion[27]和 DATFuse[28]等融合方法进行了可视化对比实验。实验选取白天、夜晚等典型场景测试,以 LLVIP 数据集为例,结果表明,本文算法在黑暗环境下表现突出,传统方法在暗区容易丢失纹理细节,本文算法能够清晰保留目标边缘和黑暗中的行人信息,同时抑制噪声。尤其在黑暗环境下表现突出,具有更强的光照适应性和泛化能力,结果如图 11 所示。

(2)定量分析

在 MSRS 和 LLVIP 数据集上测试多个代表性多光谱

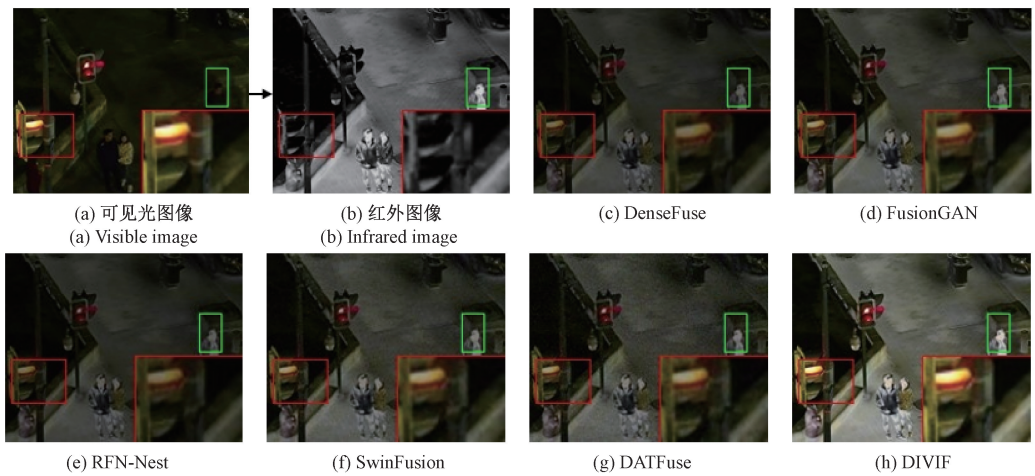

图 11 不同方法在 LLVIP 数据集上的结果

Fig. 11 Performance of different methods on the LLVIP dataset

检测网络。表 3 和 4 为数据集上多种算法模型的定量对比结果。定量分析表明,DIVIF 在所有指标上均表现最优,其在图像细节保留和清晰度方面具有显著优势。

表 3 在 MSRS 数据集上的定量对比

Table 3 Quantitative comparison on the MSRS dataset

算法模型	SD	VIF	AG	EN	Qabf	SF
DenseFuse (2019)	21. 720 5	0. 752 8	1. 867 3	5. 510 4	0. 425 6	5. 690 2
FusionGAN (2019)	17. 160 8	0. 401 7	1. 223 6	5. 267 3	0. 129 5	3. 855 7
RFN-Nest (2021)	21. 550 7	0. 589 2	1. 248 9	5. 372 1	0. 230 5	4. 365 8
SwinFusion (2022)	32. 920 5	0. 942 3	2. 690 8	5. 897 6	0. 592 4	8. 810 7
DATFuse (2023)	27. 120 5	0. 861 3	2. 630 4	5. 815 6	0. 528 9	8. 845 7
DIVIF	**63. 247 7**	**0. 908 5**	**4. 088**	**7. 128 2**	**0. 68 2**	**12. 509 3**

表 4 在 LLVIP 数据集上的定量对比

Table 4 Quantitative comparison on the LLVIP dataset

算法模型	SD	VIF	AG	EN	Qabf	SF
DenseFuse (2019)	34. 210 5	0. 752 3	2. 735 8	6. 860 4	0. 501 2	9. 250 7
FusionGAN (2019)	24. 805 3	0. 462 8	1. 932 6	6. 320 7	0. 268 4	6. 905 2
RFN-Nest (2021)	34. 630 5	0. 682 1	2. 143 6	6. 875 9	0. 328 4	6. 305 7
SwinFusion (2022)	44. 905 6	0. 818 7	3. 930 2	7. 142 3	0. 635 4	13. 610 8
DATFuse (2023)	39. 865 7	0. 823 6	3. 050 2	7. 068 3	0. 497 8	12. 350 2
DIVIF	**52. 746 5**	**0. 833 1**	**4. 835 2**	**7. 557 7**	**0. 723 8**	**16. 845 4**

在 LLVIP 和 MSRS 数据集上对 DIVIF 进行了验证,实验结果表明,DIVIF 在定性和定量评估中均优于 5 种经典算法,生成的融合图像边缘清晰、信息完整,显著提升了后续多模态行人检测的适用性。

4.2　改进式法检测算法仿真实验

1)实验参数设置

实验环境配置如下:采用 Python3.8+Pytorch1.9.0+Cuda11.1 搭建网络,使用 RTX 3060 GPU 系统,图像分辨率为 640×640,训练参数如表 5 所示。

表 5　网络训练参数

Table 5　Network training parameters

参数	指标
初始学习率	0.01
权重衰减系数	0.000 5
优化方法	AdamW
Batch Size	16
Epoch	200

实验基于 LLVIP 数据集中的图像展开,使用 LabelImg 工具对数据集进行人工标注,标注类别为 people,并按照 8∶2 的比例将数据集划分为训练集和验证集,以支持模型的训练与性能评估。

2)检测算法评价指标

本文采用 4 种定量指标综合评估目标检测模型性能:平均精度均值(mean average precision,mAP)作为核心指标,反映模型整体检测精度;模型参数量(parameters)表征网络复杂度,直接影响模型轻量化程度;浮点运算次数(floating point operations,FLOPs)衡量计算复杂度,决定模型运行效率;每秒帧数(FPS)评估实时性,反映实际应用价值;平均精度(average precision,AP)通过精确率-召回率曲线面积量化单类别检测性能;推理时间(inference time)直接体现模型处理速度,是工程应用的关键指标。这些指标从精度、效率、实用性等维度全面评估模型性能。

3)网络轻量化结果分析

(1)主干网络轻量化结果分析

基于 MobileNetV4 优化 YOLOv8n 主干网络,在 LLVIP 数据集上进行 4 项性能指标评估,具体量化结果如表 6 所示。

表 6　主干网络轻量化结果

Table 6　Lightweight backbone network results

模型	mAP/%	Parameters/($\times10^6$)	GFLOPs	FPS
YOLOv8n	93.6	3.05	8.0	121
YOLOv8n+MobileNetV4	93.5	2.10	5.6	155

表 6 中,改进后的模型在 Parameters 和 GFLOPs 均降低 30%的同时,FPS 提升 28%,而 mAP 仅下降 0.1%,基本保持检测精度。该算法显著提升了模型效率,为实际应用提供了更优的解决方案。

(2)颈部网络轻量化结果分析

将颈部网络中的 C2f 模块结合 GSConv 替换成 C2f-GSC 模块,进行综合性能分析,结果如表 7 所示。

表 7　颈部网络轻量化结果

Table 7　Lightweight neck network results

模型	mAP/%	Parameters/($\times10^6$)	GFLOPs	FPS
YOLOv8n	93.6	3.05	8.0	121
YOLOv8n+C2f-GSC	94.0	2.82	7.5	129

表 7 中,改进算法使 mAP 提升 0.4%,同时降低参数量和 GFLOPs,FPS 从 121 提升至 129。结果表明,该算法在行人检测任务中有效提升了推理速度和检测效率,同时保持了较高的检测精度。

4)添加注意力机制结果分析

在 YOLOv8n 中分别集成 SE、CBAM 和 EMA 模块构建模型,结果如表 8 所示。

表 8　添加注意力机制结果

Table 8　Attention-enhanced results

模型	mAP/%	Parameters/($\times10^6$)	GFLOPs	FPS
YOLOv8n	93.6	3.05	8.0	121
YOLOv8n+SE	93.8	3.07	8.2	115
YOLOv8n+CBAM	94.1	3.08	8.5	110
YOLOv8n+EMA	94.6	3.04	8.3	104

表 8 中,采用 EMA 注意力机制的模型在检测精度上表现最优,其 mAP 相较于 SE 和 CBAM 注意力机制分别提升了 0.8%和 0.5%,在模型效率指标上,其参数量和浮点计算数均优于对比方法。结果表明,所采用的 EMA 模块不仅有效提升了检测精度,同时保持了模型的计算效率,实现了性能与复杂度的良好平衡。

5)改进损失函数结果分析

对 YOLOv8n 框架的 CIoU 边界框回归损失替换为 MPDIoU 损失函数进行对比试验,结果如表 9 所示。

表 9　改进损失函数结果

Table 9　Enhanced loss function results

模型	mAP/%	Parameters/($\times10^6$)	GFLOPs	FPS
YOLOv8n	93.6	3.05	8.0	121
YOLOv8n+MPDIoU	94.2	3.05	8.0	120

表 9 中,在保持参数量、计算量和推理速度基本不变的情况下,mAP 值提升了 0.6%。这一结果验证了 MPDIoU 在目标检测任务中的有效性,展现了其在复杂场景下边界框回归的优越性能。

6)对比实验分析

本文系统评估了改进后的 YOLOv8n 网络,并与两类典型目标检测模型进行对比,以 YOLO 系列为代表的单阶段检测器和基于 Faster R-CNN 的双阶段检测框架,数

据如表 10 所示。

表 10 不同目标检测网络结果对比

Table 10 Comparative analysis of object detection networks

模型	Precision/%	mAP/%	Parameters/($\times10^6$)	FPS
FasterR-CNN	90.8	92.4	43.52	8
YOLOv5n	89.2	91.0	2.87	113
YOLOv7-tiny	88.1	90.2	4.20	105
YOLOv8n	92.5	93.6	3.05	121
YOLOv8n-MGEL	94.6	95.3	1.75	142

实验结果表明,Faster R-CNN 作为二阶段检测器 mAP 虽高,但参数量达到本文方法的 25 倍;YOLOv5n、YOLOv7n 等单阶段算法各具优势,但各项指标均低于 YOLOv8n;本文提出的 MGEL 改进算法,mAP 达到 95.3%,较基础 YOLOv8n 提升 1.7%。可视化实验显示,如图 12 和 13 所示,该算法在弱光/强光环境下均保持优异性能,验证了其在红外-可见光融合图像行人检测中的综合优势,在保证检测精度的同时,实现了更轻量化的网络设计。

(a) 原始融合图像
(a) Initial fusion output

(b) YOLOv8n检测效果
(b) YOLOv8n detection performance

(c) YOLOv8n-MGEL检测效果
(c) Detection performance of YOLOv8n-MGEL

图 12 低光照条件下的检测结果对比

Fig. 12 Low-illumination detection benchmark

本实验针对目标检测模型参数量大、精度不足及低光照交通场景检测难题,测试基于 YOLOv8n 的轻量化行人检测网络性能。通过 MobileNetV4 主干网络和 GSConv 颈部结构压缩参数,结合 EMA 注意力模块增强特征提取,选用 MPDIoU 损失函数提升检测准确率。实验表明,优化后的网络模型检测精度达到 95.3%,显著优于对比模型,证实了所提方法在行人检测任务中的有效性。算法的轻量化设计,使其可部署于计算资源受限的车载终端,满足智能驾驶系统对实时目标检测的需求。

(a) 原始融合图像
(a) Initial fusion output

(b) YOLOv8n检测效果
(b) YOLOv8n detection performance

(c) YOLOv8n-MGEL检测效果
(c) YOLOv8n-MGEL detection performance

图 13 高光照条件下的检测结果对比

Fig. 13 High-illumination detection benchmark

5 基于红外-可见光融合图像行人检测系统集成与应用

为了推动行人检测技术的实际应用,本研究开发了基于红外-可见光融合图像行人检测系统,集成图像融合算法和改进检测模型,实现图像预处理到目标检测全流程。

5.1 系统功能架构分析及设计

本文开发了一种基于红外与可见光图像融合的行人检测系统,采用离线训练模式直接加载预训练权重,实现高效精准的检测,结合多光谱融合技术提升检测精度。系统支持双光谱图像的上传、预处理(去噪/对齐)及特征融合,实现高质量行人检测与可视化标注存储,配备交互式界面满足实时处理、高精度、低延迟需求,可全自动备运行并支持多设备数据溯源管理。

5.2　系统测试

本研究硬件设备采用杰锐微通 SM500 双目摄像头，支持彩色/红外成像；软件基于 Python3.8+Pytorch1.9.0+YOLOv8 框架，使用 PyQt5 开发检测界面，具体界面如图 14 所示。通过在实验车上进行夜间道路测试，验证了系统在夜间低光照条件下的多光谱融合效果和行人检测性能。

图 14　场景测试结果示例

Fig. 14　Scenario testing results demonstration

实验结果表明，本系统完整实现了多光谱行人检测的全流程功能架构。系统集成了图像融合与行人识别等核心功能模块，能够准确检测并提示道路行人信息。该系统具备双模态图像采集与预处理能力，通过特征融合算法有效整合红外与可见光图像的优势特征，并基于深度学习模型实现复杂场景下的行人目标精准识别。系统提供交互式可视化界面，支持检测结果的实时展示与数据管理功能，形成完整的智能决策支持闭环。

在性能表现方面，系统展现出卓越的环境适应性和计算效率。通过多源信息融合策略，系统在各类复杂场景下均保持稳定的检测性能，处理流程满足严格的实时性要求。该系统的应用有助于提升无人驾驶和辅助驾驶车辆在复杂交通环境中的行驶安全性，从而有效减少交通事故的发生率。系统采用高度自动化的处理架构，具备良好的硬件兼容性和数据可追溯性，为智能安防和自动驾驶应用提供了可靠的技术解决方案。

6　结　论

本研究开发了一套基于红外与可见光融合的行人检测系统，通过多模态数据融合提升了复杂环境下的检测性能。系统采用离线训练模式，在实际应用中表现出良好的检测精度和运行效率，为智能安防和自动驾驶领域提供了有效的技术解决方案。研究完成了算法突破和系统实践，并在实际交通场景中验证了技术的可行性。未来将重点突破动态场景适应性和边缘计算部署等关键技术难题，进一步推动该技术在智能交通等领域的实际应用。

参考文献

[1] 别倩，王晓，徐新，等. 红外—可见光跨模态的行人检测综述[J]. 中国图象图形学报，2023，28(5)：1287-1307.
BIE Q，WANG X，XU X，et al. Review of infrared-visible cross-modal pedestrian detection[J]. Journal of Image and Graphics，2023，28(5)：1287-1307.

[2] 谢富，朱定局. 深度学习目标检测方法综述[J]. 计算机系统应用，2022，31(2)：1-12.
XIE F，ZHU D J. Review of deep learning-based object detection methods [J]. Computer Systems & Applications，2022，31(2)：1-12.

[3] VIOLA P，JONES M. Rapid object detection using a boosted cascade of simple features[J]. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition，2001，23 (9)：511-518.

[4] GIRSHICK R，DONAHUE J，DARRELL T，et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2014：580-587.

[5] 蔡腾，陈慈发，董方敏. 结合 Transformer 和动态特征融合的低照度目标检测[J]. 计算机工程与应用，2024，60 (9)：135-141.
CAI T，CHEN C F，DONG F M. Low-illumination object detection combining Transformer and dynamic feature fusion[J]. Computer Engineering and Applications，2024，60(9)：135-141.

[6] 王满利，张航，张长森. 基于深度学习的低光照目标检测算法[J]. 北京邮电大学学报，2024，47(5)：59-65.
WANG M L，ZHANG H，ZHANG CH S. Low-illumination object detection algorithm based on deep learning[J]. Journal of Beijing University of Posts and Telecommunications，2024，47(5)：59-65.

[7] 李宝兵，符长友. 一种改进 YOLOv8n 的低光照行人检测算法[J]. 西华大学学报(自然科学版)，2025，26(6)：826-833.
LI B B，FU CH Y. An improved YOLOv8n low-illumination pedestrian detection algorithm[J]. Journal of Xihua University (Natural Science Edition)，2025，26(6)：826-833.

[8] 赵京楠，刘洋，白雪琼. AF-YOLOv8：双光自适应融合行人检测算法[J]. 红外与激光工程，2025，54(10)：

123-131.
ZHAO J N, LIU Y, BAI X Q. AF-YOLOv8: Dual-optical adaptive fusion pedestrian detection algorithm [J]. Infrared and Laser Engineering, 2025, 54(10): 123-131.

[9] 姜柏军, 钟明霞, 林昊昀. 深度学习与图像融合的行人检测算法研究[J]. 激光与红外, 2024, 54(2): 302-306.
JIANG B J, ZHONG M X, LIN H Y. Research on pedestrian detection algorithm based on deep learning and image fusion [J]. Laser & Infrared, 2024, 54(2): 302-306.

[10] 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.
LIU F, WANG S B, WANG X J. Multi-feature cascade infrared pedestrian detection method in low visibility environment[J]. Infrared and Laser Engineering, 2018, 47(6): 137-144.

[11] LUPIÓN M, POLO-RODRÍGUEZ A, ORTIGOSA P M, et al. Thermal YOLO: A person detection neural network in thermal images for smart environments [C]. Proceedings of the International Conference on Ubiquitous Computing & Ambient Intelligence.

[12] 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(9): 1070-1080.
DAI SH SH, LIU K SH, HUANG L, et al. Infrared small target detection method based on visual transformer and dual decoder [J]. Infrared Technology, 2024, 46(9): 1070-1080.

[13] 段沛沛, 张严, 雒明世, 等. 基于 U 型多尺度 Transformer 网络的红外小目标检测算法[J]. 西北工业大学学报, 2025, 43(1): 154-161.
DUAN P P, ZHANG Y, LUO M SH, et al. Infrared small target detection algorithm based on U-shaped multi-scale transformer network [J]. Journal of Northwestern Polytechnical University, 2025, 43(1): 154-161.

[14] 张上, 崔玉杰, 龚国强. LIST-DETR: 复杂背景下红外小目标检测的高效轻量级检测网络[J]. 红外与激光工程, 2025, 54(10): 201-209.
ZHANG SH, CUI Y J, GONG G Q. LIST-DETR: An efficient lightweight detection network for infrared small target detection in complex backgrounds[J]. Infrared and Laser Engineering, 2025, 54(10): 201-209.

[15] 刘团囡,刘淑娴,哈妮克孜·伊拉洪,等. Trans-YOLO: 基于 RT-DETR 解码器改进 YOLOv8 的红外小目标检测[J]. 激光与光电子学进展,2025,62(10):378-389.
LIU J N, LIU SH X, YILLAHONG H, et al. Trans-YOLO: Infrared small target detection by improving YOLOv8 with RT-DETR decoder [J]. Laser & Optoelectronics Progress,2025,62(10):378-389.

[16] 李朝旭,徐清宇,安玮,等. 红外图像暗弱目标轻量级检测网络[J]. 红外与毫米波学报, 2025, 44(2): 300-310.
LI ZH X, XU Q Y, AN W, et al. Lightweight detection network for dim and weak targets in infrared images[J]. Journal of Infrared and Millimeter Waves, 2025, 44(2): 300-310.

[17] 于洋. 基于大模型的跨模态对齐方法研究基于强化联邦学习和大模型赋能的突发事件跨模态搜索与态势预测[D]. 北京邮电大学, 2025: 987-1002.
YU Y. Research on cross-modal alignment method based on large models: Reinforced federated learning and large model empowered cross-modal search and situation prediction for emergencies [D]. Beijing University of Posts and Telecommunications, 2025: 987-1002.

[18] 戴佳惠,黄敏,肖仲喆. 结合迁移学习和 Transformer 模型的低资源多模态语音情感识别[J]. 声学技术, 2025, 44(5): 722-729.
DAI J H, HUANG M, XIAO ZH ZH. Low-resource multimodal speech emotion recognition combining transfer learning and transformer model[J]. Technical Acoustics, 2025, 44(5): 722-729.

[19] 易小西,张 驰. 基于多模态信息源融合的目标检测策略研究[J]. 自 动 化 应 用, 2025, 66(18): 99-108.
YI X X, ZHANG CH. Research on target detection strategy based on multimodal information source fusion [J]. Automation Application, 2025, 66(18): 99-108.

[20] YANG C, WANG Y, ZHANG J, et al. A. Lite vision transformer with enhanced self-attention[J]. Journal of Computer Vision and Pattern Recognition, 2022, 45(3): 1234-1245.

[21] QIN D, LEICHNER C, DELAKIS M, et al. MobileNetV4: Universal models for the mobile ecosystem[C]. European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 78-96.

[22] LI H, LI J, WEI H, et al. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles [J]. arXiv preprint arXiv, 2022: 2206.02424.

[23] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module for feedforward convolutional neural networks [C]. Proceedings of the

European Conference on Computer Vision (ECCV). Munich: Springer, 2018: 3-19.

[24] LI H, WU X J. DenseFuse: A fusion approach to infrared and visible images[J]. IEEE Transactions on Image Processing, 2019, 28(5): 2614-2623.

[25] MA J Y, YU W, LIANG P W, et al. FusionGAN: A generative adversarial network for infrared and visible image fusion[J]. Information Fusion, 2019, 48: 11-26.

[26] 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.

[27] 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.

[28] TANG W, HE F Z, LIU Y, et al. DATFuse: Infrared and visible image fusion via dual attention transformer[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2023, 33(7): 3159-3173.

作者简介

张立成,2009 年于长安大学获得学士学位,2012 年于长安大学获得硕士学位,2022 年于长安大学获得博士学位,现为长安大学高级工程师,主要研究方向为交通图像处理、汽车能耗预测模型、汽车测试技术与应用。

E-mail: lichengzhang@chd.edu.cn

Zhang Licheng received his B. Sc. degree from Chang′an University in 2009, M. Sc. degree from Chang′an University in 2012, and Ph. D. degree from Chang′an University in 2022, respectively. Now he is a senior engineer in Chang′an University. His main research interests include traffic image processing, vehicle energy consumption prediction models, automobile testing technology and applications.

王蕾(通信作者),2005 年于西安工业大学获得学士学位,2010 年于长安大学获得硕士学位,现为空军工程大学副教授,主要研究方向为人工智能、图像与目标识别。

E-mail: xiaoci2000@sina.com

Wang Lei (Corresponding author) received her B. Sc. degree from Xi′an Technological University in 2005, and M. Sc. degree from Chang′an University in 2010, respectively. Now she is an associate professor at Air Force Engineering University. Her main research interests include signal and information processing.

宋逸菲,2022 年于西安邮电大学获得学士学位,2025 年于长安大学获得硕士学位,主要研究方向为红外可见光图像融合和深度学习。

E-mail: 13093963732@163.com

Song Yifei received her B. Sc. degree from Xi′an University of Posts and Telecommunications in 2022, and M. Sc. degree from Chang′an University in 2025. Her main research interests include infrared visible image fusion and deep learning.

谢耀华,2007 年于长安大学获得学士学位,2010 年于长安大学获得硕士学位,现为招商局重庆交通科研设计院有限公司研究员/正高级工程师,主要研究方向智能交通安全与管控。

E-mail: 286262813@qq.com

Xie Yaohua received his B. Sc. degree from Chang′an University in 2007 and M. Sc. degree from Chang′an University in 2010. Now he is a professor/senior engineer at China Merchants Chongqing Communications Technology Research & Design Institute Co. Ltd. His main research interests include intelligent traffic safety and control.

周竹萍,2005 年于东南大学获得学士学位,2008 年于东南大学获得硕士学位,2012 年于东南大学获得博士学位,2020 年当选江苏省科技人才项目副主席。现为南京理工大学自动化学院教授,研究方向包括智能交通系统、交通视频检测技术等。

Zhou Zhuping received her B. Sc. degree from Southeast University in 2005, M. Sc. degree from Southeast University in 2008, and Ph. D. degree from Southeast University in 2012. Now she is a professor at the School of Automation, Nanjing University of Science and Technology. Her main research interests include intelligent transportation systems, and traffic video detection technologies.

陈春成,2023 年于西安建筑科技大学获得学士学位,长安大学在读研究生。主要研究方向为图像处理和深度学习。

E-mail: 2023224111@chd.edu.cn

Chen Chuncheng received his B. Sc. degree from Xi′an University of Architecture and Technology in 2023. Now he is a M. Sc. candidate at Chang′an University. His main research interests include image processing and deep learning.