DOI: 10.19650/j.cnki.cjsi.J2413224

基于梯度算子和注意力的多模态融合目标检测*

李学钊,王 伟,薛 冰
(哈尔滨工程大学智能科学与工程学院 哈尔滨 150000)

摘 要:红外与可见光图像具有很好的互补特性,可以利用这 2 种模态图像的融合来适应自动驾驶等领域对于目标检测高精度和高鲁棒性的要求。现有多模态目标检测算法往往模型庞大,推理耗时长,无法在边缘设备上部署,而采用直接融合等方法又无法充分发挥不同模态的优势,因此提出了一种基于梯度算子和注意力机制的融合目标检测算法。引入梯度算子设计定制化卷积来捕获图像纹理;红外支路引入坐标注意力发挥其目标定位优势;引入权重生成网络对 2 个模态的特征进行自适应加权融合。算法结构模块化,轻量化,适合部署在边缘设备上。在数据集上实验,得到 mAP@0.50 和 mAP@0.5∶0.95 指标值比可见光单模态检测提升了 6.3%和 7.2%,比红外提升了 11.3%和 9.8%。推理帧率可达 22.7,满足实时性要求。

关键词: 目标检测;双模态;特征融合;梯度算子;注意力机制
中图分类号: TH741 TP391.41 **文献标识码**: A **国家标准学科分类代码**: 510.4050

Multi-modal fusion object detection based on gradient operator and attention

Li Xuezhao, Wang Wei, Xue Bing
(*School of Intelligent Science and Engineering, Harbin Engineering University, Harbin 150000, China*)

Abstract: Infrared and visible images exhibit complementary characteristics, making their fusion highly suitable for achieving high accuracy and robustness in target detection for applications such as autonomous driving. However, existing multimodal object detection algorithms often feature large models and long inference times, making them unsuitable for deployment on edge devices. Additionally, direct fusion methods fail to fully leverage the strengths of different modalities. To address these challenges, we propose a fusion object detection algorithm that integrates a gradient operator and an attention mechanism. A gradient operator is employed to design a customized convolutional layer for capturing image texture. In the infrared branch, coordinate attention is incorporated to enhance target localization capabilities. Additionally, a weight generation network is introduced to adaptively balance the features of both modalities. The algorithm is modular and lightweight, making it ideal for edge device deployment. Experiments on benchmark datasets demonstrate that the proposed method achieves mAP@0.50 and mAP@0.5∶0.95 scores that are 6.3% and 7.2% higher, respectively, than single-modal detection using visible images, and 11.3% and 9.8% higher than infrared detection. The inference frame rate reaches 22.7 FPS, meeting real-time processing requirements.

Keywords: object detection; dual-modal; feature fusion; gradient operator; attention mechanism

0 引 言

在现代计算机视觉领域,目标检测作为核心任务之一,广泛应用于自动驾驶[1]、监控系统[2]、智能安防等各个领域。从目前的研究趋势可以看出,深度学习已经成为解决目标检测任务的主流方法,并且通常分为两阶段检测和单阶段检测。两阶段检测器首先生成候选区域(region proposals),然后对这些候选区域进行分类和边界框回归,通常具有较高的检测精度。典型代表包括:包括 R-CNN[3]、Fast R-CNN[4]、Faster R-CNN[5]和 Mask R-CNN[6]等。单阶段检测网络在单次前向传递过程中同时生成边界框和类别预测,主要包括 SSD[7]网络和 YOLO 系列[8-11]。这些方法在图像的不同尺度上生成候选框,由于设计更简单,它们的实时性能显著优于两阶段检测网络。尤其是 YOLO v5s 实现了精度和速度

收稿日期:2024-08-28 Received Date: 2024-08-28
*基金项目:江淮前沿技术协同创新中心追梦基金课题(2023ZM01Z025)项目资助

的出色平衡,因此本文以此方法作为检测部分的基础框架。

目标检测任务通常以各种传感器作为原始输入,传统的目标检测方法主要依赖于可见光图像。可见光图像具有丰富的纹理和颜色信息[12],在正常光照条件下这些方法表现出色。然而,在强光照射、夜间、遮挡严重、以及隧道等光线突然变化的恶劣环境下,可见光图像的有效性显著降低。在实际应用中,其鲁棒性无法达到值得信赖的水平。相比之下,红外传感器由于其能够在无光照或低光照条件下捕捉到物体的热辐射信息,在上述恶劣环境中表现良好[13]。然而,红外传感器采集的图像可用的特征较少,并且受温度影响很大,单一红外图像亦有其局限性。由上可以看到,红外图像与可见光图像具有很好的互补特性,因此融合红外和可见光双模态信息的目标检测方法成为提升目标检测精度和鲁棒性的一个重要研究方向。Deng 等[14]提出了一种多层融合网络,从每个主干块的 RGB 通道和红外通道创建多尺度特征图,并引入特征金字塔模块,提高了在低光环境下的检测效果。Wu 等[15]提出了一种基于形态特征、红外辐射和运动特征的多模态特征融合网络,以弥补单一模态特征描述小目标的不足。为了解决跨模态特征建模的挑战,Zhao 等[16]提出了一种相关性驱动的特征分解融合网络,并且提出了一种相关驱动损失,促使两分支提取过程中低频特征相关而高频特征不相关,在统一的基准测试中提高了下游目标检测的性能。这些检测方法广泛提高了恶劣环境下目标检测的稳定性。

根据融合层次不同,现有的多模态融合方法可分为像素级融合、特征级融合和决策级融合。像素级融合又被称为前期融合,它直接对图像中的像素信息进行融合,更好地保留了源图像的有效信息。然而,像素级融合面临不同模态之间的对齐和配准问题,且融合后数据量大,处理复杂。决策级图像融合将不同源图像处理部分作出的决策进行融合,得到最终结果。这种方法各模态的处理过程独立,互不影响,可以充分利用各自的优势,然而各模态单独处理可能存在冗余计算,并且决策层的融合策略设计复杂。特征级融合又称中期融合,分别从不同模态的图像中提取特征,然后将这些特征进行融合。其方法更灵活,针对图像中隐含的高维特征, 可以采用多种特征提取和融合策略。

以主流检测模型 YOLOv5-s 为基本框架, 对其主干特征提取网络进行改进。仅改变原 CSPDarknet-53 网络的一层 Conv 和 C3 模块就可以实现对跨模态输入的建模,避免了引入复杂的结构使模型推理速度降低,保证实时性;通过结构设计充分利用 2 种模态互补的信息并降低了人工设置融合规则的不合理风险。提高了检测精度和鲁棒性。

1　融合红外-可见光信息的目标检测方法

基于 YOLOv5-s 的检测网络,本文设计了一个基于梯度算子和注意力机制的红外-可见光多模态融合检测网络,如图 1 所示,本文作为对比的基准模型 Baseline 的主干如图 1(a)所示,借鉴文献[17]利用 2 个相同 CSPDarknet53 构成双分支网络结构,融合方式为特征图直接相加。整体框架如图 1(b)所示,其中,Conv 部分均为延续 YOLOv5-s 中卷积、批归一化和激活函数的系列操作。本文主要针对跨模态特征建模问题,改进特征提取主干网络 CSPDarknet53,让颈部部分以及检测器能够充分理解并利用红外和可见光 2 种模态互补的信息,从而应对更加全面的场景。

本文算法设计了梯度融合注意力模块(gradient fusion attention module, GFAM)和双模态特征融合模块(dual-modal feature fusion module, DFFM)的组合,前者着重于充分提取 2 种模态各自有关目标信息的特征,后者着重于利用卷积层结构自适应的学习融合规则,以充分理解与融合目标信息。

1.1　改进的特征提取主干

改进的特征提取主干网络可以接收红外与可见光 2 种模态的图像 $\{I,V\}$ 作为输入,在特征提取的初步阶段,经过 2 个卷积层进行浅层特征提取后得到两张不同模态的特征图 $\{F_{IR},F_{VI}\}$, 之后设计了 GFAM 模块,通过对两分支梯度信息的交互以及针对各自模态特点的特征提取,得到具有模态特点的特征图 $\{\Phi_I^G,\Phi_V^G\}$。再将该特征图通过 DFFM 模块得到融合特征图 Φ_F^D 后经过 YOLOv5-s 原模型的后续网络,进行目标分类和定位。得到融合特征图 Φ_F^D 的过程可以描述为:

$$F_{IR}=f_2(I),F_{VI}=f_2(V) \tag{1}$$

$$\Phi_I^G=G(F_{IR}),\Phi_V^G=G(F_{VI}) \tag{2}$$

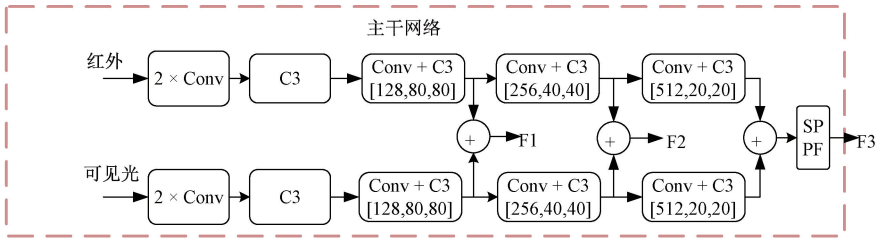
$$\Phi_F^D=D(\Phi_I^G+\Phi_V^G) \tag{3}$$

其中,$f_i(\cdot)$ 表示 i 层 Conv 的作用, $G(\cdot)$ 表示 GFAM 模块作用, $D(\cdot)$ 表示 DFFM 模块的作用。

1.2　梯度融合注意力模块

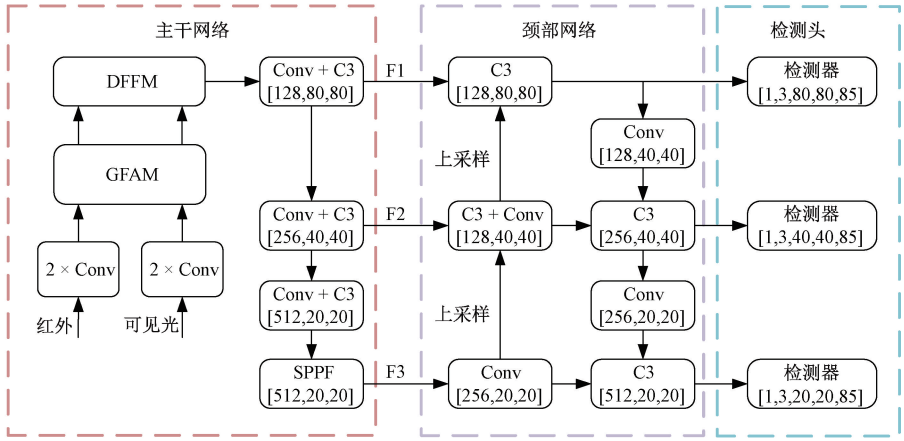
目前,常见的跨模态图像融合的方法一般都是采用双分支独立的特征提取网络,然后通过不同的方式将两者的结果进行融合。在该过程中,两分支特征提取过程没有相互指导,往往会忽略模态之间共享的特征[17]。就红外与可见光图像不同模态的成像特点,设计了如图 2 所示的梯度融合注意力模块,替换 YOLOv5-s 单路特征提取单元 C3 模块,在有限增加模型结构的前提下,优化模型之间的交互。

虽然可见光相比于红外图像拥有更加丰富的纹理



(a) 直接融合方法Baseline的主干

(a) The Backbone of the direct fusion method Baseline



(b) 改进模型

(b) Improving the model

图 1 红外-可见光融合检测网络框架

Fig. 1 Framework of infrared-visible light fusion detection network

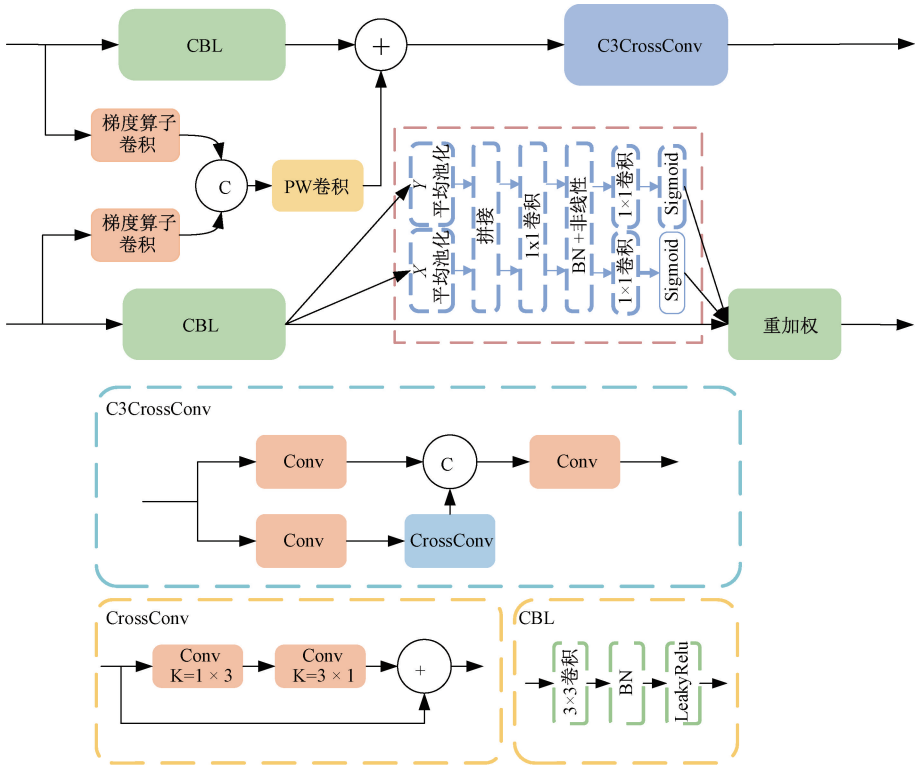


图 2 梯度融合注意力模块

Fig. 2 Gradient fusion attention module

细节信息,但有研究对比发现红外也同样拥有部分纹理信息[18]。通过将红外图像的纹理特征提取出来与可见光图像的特征进行交互,可以有效增加网络对目标的检测精度。借鉴 Resblock[19]的设计,模块首先设计了一个梯度融合交互的类残差结构,其中主流部分采用普通的 CBL 卷积组合,包括 3×3 的卷积层、正则化 BN 层和 LeakyRelu 激活函数,残差部分采用两路梯度算子卷积以及一个逐点卷积(point-wise conv,PW)。梯度算子卷积是手动设计卷积核的特殊卷积操作,可以通过引入梯度算子将输入特征与高频卷积核进行卷积,来获取图像的梯度信息,捕获物体的纹理细节。该过程可以表示为:

$$F'_{VI} = f_1(F_{VI}) \oplus PWConv(C(\nabla F_{IR} + \nabla F_{VI})) \tag{4}$$

其中,F'_{VI} 是输出的可见光特征图,∇是指梯度算子,$PWConv(\cdot)$ 代表逐点卷积操作,$C(\cdot)$ 代表 Concat 拼接操作,$\oplus$表示逐元素求和。该部分用梯度幅度信息聚合可学习的卷积特征。

在本文研究中,此处利用 Sobel 算子来计算梯度大小,所使用的 Sobel 梯度算子为:

$$\begin{cases} G_X = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \times F \\ G_Y = \begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} \times F \end{cases} \tag{5}$$

其中,G_X 和 G_Y 分别为水平和垂直方向上从特征图 F 中获取的梯度信息。

PW 卷积使用 1×1 大小的卷积核,通常不改变特征图大小而改变通道数,可以在深度上对输入特征图进行融合。与标准卷积相比,PW 卷积的计算量和参数量都相对较小。本文利用 PW 卷积对拼接后的特征图进行降维处理,使其可以与主流部分的输出维度一致。

经过梯度的融合交互之后,根据红外与可见光图像各自的特点,采用不同的方式来提取多尺度目标特征。目标检测网络的 Backbone 部分应该尽量的去提取图像中目标位置的特征信息,为了发挥红外图像[20]的目标定位优势,红外支路引入坐标注意力(coordinate attention,CA)机制。注意力机制源于对人类视觉研究,强调在关注目标或场景时[21]对不同空间部分分配不同的注意力,在捕获整个图像中的焦点区域方面发挥着重要的作用[22]。CA 通过引入空间坐标信息来增强特征表示,能有效地捕捉特征图中重要的空间位置和通道间的依赖关系,在不显著增加计算复杂度的情况下,提高了模型的定位检测能力。可见光支路采用了由交叉卷积(cross convolution,CrossCor)组成的 C3CrossConv 模块用于特征提取和多层次融合,包括 2 个标准卷积层,它们在特征图上以交叉模式排列。与传统的 k×k 大小的滑动窗口卷积不同,第 1 层交叉卷积使用 1×k 大小的内核,水平步长为 1,垂直步长为 s,第 2 层使用 k×1 内核,在 2 个维度上步长均为 s。C3CrossConv 和 CrossConv 的结构如图 2 的下方所示,这样的结构可以减少模型的参数且使模型更加紧凑。

1.3　双模态特征融合模块

可见光图像中包含着丰富的纹理细节信息,红外图像中包含显著的对比度及目标定位信息,2 种模态特征的结合能够实现跨模态信息互补,提高检测的精度和稳定性。要使得 YOLOv5 的颈部和检测头能同时学习到红外与可见光提取出来特征,就要设计一个能融合双模态特征的模块。本文研究中设计的双模态特征融合模块借鉴于注意力特征融合网络[23](attentional feature fusion,AFF),主要利用注意力机制来自适应的融合 2 个模态的特征,降低了人工设置融合规则的不合理风险。AFF 网络引入了多尺度通道注意力模块,通过 PW 卷积来关注通道的尺度问题,可以同时强调分布更广泛的大对象和分布更局部的小对象,其结构也可以保证网络能尽可能的轻量化。

本文设计的 DFFM 网络区别于 AFF 多尺度的融合任务,引入权重生成模块(weight generation module,WGM),着重解决多模态的特征融合问题,其结构如图 3 所示。

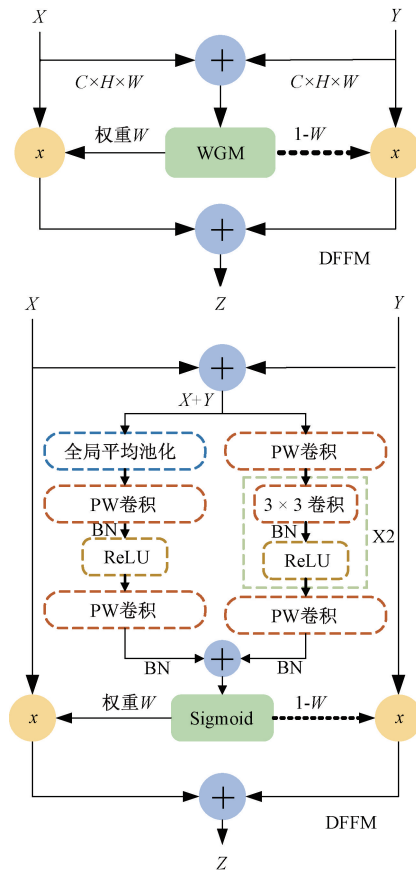


图 3　双模态特征融合模块

Fig. 3　Dual-modal feature fusion module

模型首先对输入 2 种模态的特征 X、Y 进行初步的逐元素相加融合,然后 WGM 模块使其分别经过两路不同维度的处理后求和,再经过 Sigmoid 激活函数使得输出值为 0~1 之间,得到特征 X 的融合权重值 W。这里融合策略是要让两路模态特征进行自适应加权平均,所以特征 Y 的权重就是 $1-W$。依靠模型参数在反向传播时的学习能力来自适应调整权重 W 和 $1-W$。权重与特征 X、Y 分别相乘再逐元素相加便得到了融合特征 Z。

其中,对初步的融合特征 $X+Y$ 进行左侧通道维度的重要性分析。首先经过全局平均池化,将 $C\times H\times W$ 大小特征图压缩为 $C\times 1\times 1$,即先忽略空间差异,利用 2 个轻量化的 PW 卷积来学习通道维度的注意力信息。右侧进行空间维度的注意力分析,结构上参考 BAM 模块,利用 PW 卷积对输入特征图进行降维操作,然后利用 2 个 3×3 卷积核大小的空洞卷积提取特征信息。空洞卷积又称膨胀卷积,能够在不增加参数和计算复杂度的情况下,扩大卷积核的感受野[24]。最后,利用 PW 卷积将特征图映射到 $1\times H\times W$,利用广播机制让两侧的注意力信息可以进行加和。整体上可以视为 2 种注意力网络采用并联的形式进行组合。

2 实验与结果分析

2.1 实验环境配置及参数设置

算法基于 Pytorch 实现,操作系统为 ubuntu 22.04,GPU 型号为 4090。使用随机梯度下降优化器(stochastic gradient descent, SGD)训练,动量因子为 0.937,权重衰减因子为 0.000 5。初始化学习率为 0.01,训练时输入的图片统一 resize 成 640×640 大小,batch_size 大小为 32。

2.2 数据集与评价指标

使用公开的 M^3FD[22] 多光谱目标检测数据集,其中包括 4 200 对经过校准对齐后的红外-可见光数据,该数据集涵盖了具有各种环境,照明,季节和天气的 4 个主要场景,具有广泛的像素变化范围。其中大部分图像的分辨率为 1 024×768,提供了行人、汽车、公交、摩托车、路灯和卡车 6 个类别的注释。其中,按照 8∶1∶1的比例设置训练、验证和测试集。

实验使用平均精度均值(mean average precision, mAP)和每秒帧数作为主要的评价指标。其中 mAP 值是衡量目标检测算法性能的常用指标之一,它是把查准率(precision)和查全率(recall)进行综合考虑后的值。同时,每秒帧数可以反映出算法的检测速度。

2.3 算法对比实验及分析

为了体现双模态融合检测的优势,同时验证红外与可见光特征的互补性,设计了和单模态目标检测模型 YOLOv5-s 的对比实验,包括红外和可见光图像数据集 2 种输入。为了全面评估所提算法的有效性,在 M^3FD 数据集上同当前主流的双模态融合目标检测算法进行了比较。为确保实验的一致性,在相同的硬件和软件环境下分别部署了本文算法和对比算法,比较结果如表 1 所示。从表中的数据可以观察到,在 M^3FD 数据集上,所提出的融合红外-可见光双模态信息的目标检测网络在 mAP@0.50 和 mAP@0.5∶0.95 2 个指标上的获得了 0.874、0.576 的检测结果。结果表明,提出的双模态目标检测网络在检测性能上均得到了显著提升,比经典的可见光单模态检测算法分别提升了 6.3% 和 7.2%,红外单模态分别提升了 11.3% 和 9.8%。且相比经典的双模态融合算法,如 TarDAL[25]、U2F[26]、CDDFuse[16]等,该算法在各类目标的检测精度上都展现出了明显的提升。这些结果都进一步突显了本文结构的设计在保证轻量的同时具有很高的检测精度,证明了各个模块的作用。

表 1 在 M^3FD 数据集上的对比实验

Table 1 Comparative experiments on the M^3FD dataset

模型种类	People	Car	Bus	Motor	Lamp	Truck	mAP@0.50	mAP@0.5∶0.95
YOLOv5s (可见光)	0.681	0.905	0.880	0.839	0.728	0.826	0.810	0.505
YOLOv5s (红外)	0.790	0.871	0.853	0.649	0.641	0.765	0.761	0.478
U2F[26]	0.783	0.906	0.869	0.698	0.790	0.815	0.810	0.527
TarDAL[25]	0.803	0.916	0.854	0.698	0.779	0.823	0.812	0.527
CDDFuse[16]	0.781	0.915	0.876	0.700	0.792	0.792	0.809	0.533
本文算法	**0.845**	**0.926**	**0.930**	**0.874**	**0.816**	**0.854**	**0.874**	**0.576**

2.4 消融实验

为检验本文提出的各个模块对改善双模态目标检测模型精度的效果,把它们分为 Soble 梯度算子组成的融合交互网络、C3CrossConv、CA 注意力机制和

DFFM4 种融合策略进行了消融实验,并对比了单模态和双模态 2 种输入方式的结果,其中双模态输入网络的基准为借鉴文献[17]中的 Baseline 设计,结果如表 2 所示。

表 2　在 M³FD 数据集上的消融实验

Table 2　Ablation experiments on the M³FD dataset

输入形式	提升策略				mAP@0.50	mAP@0.5:0.95	推理时间/ms
	Sobel	C3CrossConv	CA	DFFM			
可见光					0.811	0.504	0.8
红外					0.761	0.478	0.8
可见光+红外(Baseline)					0.850	0.558	1.4
可见光+红外	√				0.862	0.559	1.2
可见光+红外	√	√			0.865	0.565	1.4
可见光+红外	√	√	√		0.868	0.570	1.4
可见光+红外	√	√	√	√	**0.874**	**0.576**	1.5

模型的各个模块均对网络的检测精度有明显提升,并且均为轻量化的模块,不会显著增加模型的推理时间。由消融实验结果可以看到,虽然有时随着融合策略的使用,算法的速度稍微有所下降,但在检测精度上有显著提高。

其中,特别是 Sobel 梯度算子卷积构成的类残差结构由于其对于梯度信息的敏感性,精准提取到目标相关的信息,在计算量对比 Baseline 基准模型降低的同时,精度还提升了 1.2%。C3CrossConv 对包含梯度信息的特征图进行进一步的特征提取,效果体现在精度值的提升上。CA 结构在几乎不增加推理时间的同时,mAP@0.5:0.95 指标获得提高,符合其轻量化的设计预期。最后 DFFM 模块的加入,模型融合的规则由特征图直接相加变成可学习的自适应的融合方式,减少了融合规则由于人为设置而产生的缺陷,最终得到精度最高的实验结果。

2.5　定性分析

为了更加直观地对比单模态算法和本文所提出方法在检测任务中的表现,在数据集上进行了定性分析,结果如图 4 所示,图片由上至下分别是检测真值、可见光单模态检测结构、红外单模态检测结果和本文算法的检测结果,从左到右分别包含强灯光、夜间、遮挡和隧道 4 种对目标识别与检测不友好的场景。椭圆形虚线框代表单模态检测结果的缺陷。

由图 4 可以看到,对于第 1 列傍晚有较强车灯照射时可见光成像的效果会大大降低,图中漏检部分即为人眼无法分辨的情形;第 2 列夜晚加上行人车辆密集的场景,可见光图像中部分目标与环境无法区分,而红外图像在远处密集小目标的情况下由于纹理细节信息的缺失,目标边缘模糊,二者均出现了漏检;第 3 列代表包括浓烟、雾天等有遮挡场景,此时对目标的检测主要依赖于红外图像。第 4 列光线突然变化的隧道等场景依赖模型快速的推理能力,可见光图像在光线充足时起主导作用,光线骤降时 DFFM 模块中生成的红外权重增加。

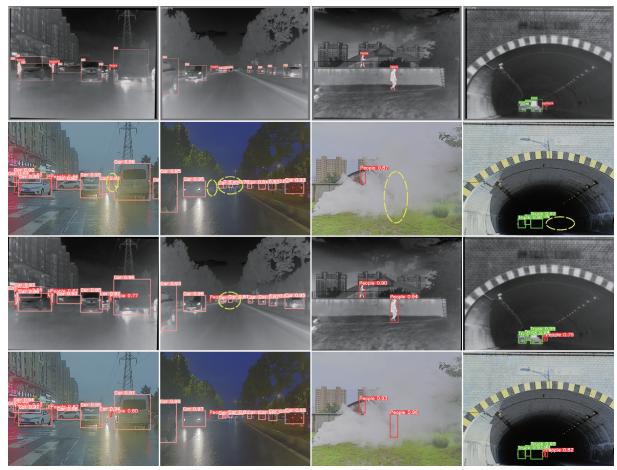


图 4　算法定性分析结果

Fig. 4　Qualitative analysis results

图 4 中椭圆框代表漏检,可见光图像在上述 4 种场景中均有漏检,这体现了可见光对恶劣环境适应能力差的缺陷;红外图像在遮挡、光照条件较差的场景中表现优秀,但是由于缺失部分纹理信息,导致在目标密集和小目标的检测中也存在漏检行为。本文算法充分考虑 2 种模态的图像信息,在各种场景的实验中均没有漏检问题,与检测的标签值基本一致,并且置信度要明显高于单一模态检测。这些定性分析结果进一步验证了红外与可见光图像良好的互补性能。

2.6　边缘设备可行性分析

为了全面分析模型的实时性能和在边缘设备上的部署的可行性,进行表 3 实验对比了本文算法的计算复杂

度,可以看到从参数量和计算量上本文的算法相比于 Baseline 有明显的优势,远小于 YOLOv5m 模型,并且参数量上跟 YOLOv5s 模型相比几乎相同,所以其在边缘设备上部署具有理论可行性。

表 3 计算复杂度及实时性实验结果

Table 3 Computational complexity and real-time experimental results

模型	参数量/M	FLOPs/G	mAP@0.50	帧率
YOLOv5s(红外)	7.02	15.8	0.78	34.1
YOLOv5m	20.9	30.1	-	-
Baseline	10.5	25.6	0.80	24.3
本文算法	7.21	21.2	0.83	22.7

为了进一步实际验证算法的实时性和应用性,设计了如图 5 所示红外和可见光 2 种模态输入的目标检测系统功能盒。整个系统接收观瞄、摄像头等输出的两路 SDI 视频作为原始输入并通过 SDI 集成式均衡解串芯片将其解串为 MIPI 信号传入核心处理板,经过算法推理后通过以太网利用实时流传输协议(real time streaming protocol,RTSP)对附有检测框的检测结果进行视频推流,作为系统的输出。其中,采用 Rockchip 发布的 RK3588 嵌入式开发板作为验证平台,RK3588 搭载八核 64 位的 ARM 架构,内置 AI 加速器 NPU,可提供 6TOPs 的算力,支持 INT4/INT8/INT16 混合运算,可以满足广泛的 AI 应用场景。

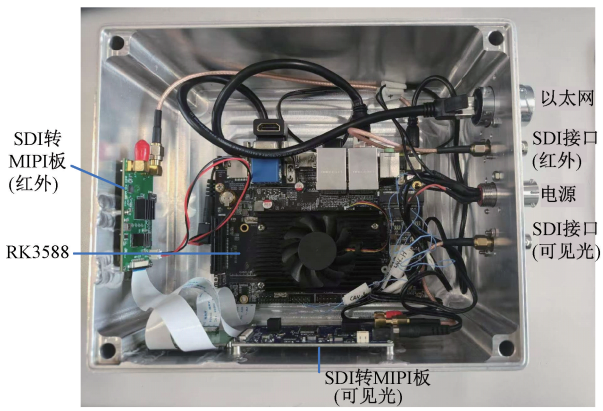


图 5 目标检测系统内部

Fig. 5 Internal view of the object detection system

将本文与对比算法的模型应用于 RK3588 平台,首先需要把 PyTorch 框架下的模型权重转换为 ONNX 兼容格式,然后利用官方 rknn-toolkit 工具对模型进行处理将其转化为支持 NPU 加速的 RKNN 模型。这个过程中会损失一部分精度,可以通过非对称或混合量化、交叉编译等操作来降低损失和提高速度。而由于 RK3588 的 NPU 加速能力能够支持的框架和算子有限,之前的图像融合网络往往包含一些 NPU 不支持的操作或层结构(例如 CDDFuse[16] 中特定类型的自注意力层、其他模型中特殊的跨尺度融合操作等),需要在 CPU/GPU 和 NPU 之间混合计算,导致推理速度较慢。将上文表 1 中经典的图像融合算法[25-27]做对比,检测框架使用 YOLOv5 网络,使用与本文算法相同的部署策略,帧率低于 1FPS,达不到实时检测的要求,无法在边缘设备上有效的部署。本文算法的推理精度和帧率具体结果如表 3 所示。从结果可以看到,部署到边缘设备上仍然可以保持高精度与较快推理速度的优势,其平均帧率可达 22.7,可以做到实时检测。

3 结 论

本文提出了一种融合红外-可见光多模态信息的目标检测方法,有效结合了红外和可见光图像特征互补的优势,在白天、夜间以及遮挡等条件下都能较准确地检测到目标。在 M^3FD 数据集上,与单模态的检测以及与多个经典的双模态目标检测算法进行了对比,结果都验证了本文的方法能够显著提升检测的性能,同时具有良好的实时性。为了更加直观地对比单模态算法和本文所提出方法在检测任务中的表现,在数据集上进行了定性分析,其中涵盖了强光、夜晚、遮挡等复杂环境,验证了本文算法具有很好的鲁棒性。最后设计了一套多模态目标检测的功能盒,在 rk3588 平台上部署本文算法,验证了算法的实时性能。

参考文献

[1] LU AN D, QIAN C, LI CH L, et al. Duality-gated mutual condition network for RGBT tracking[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022: 1-14.

[2] LI CH L, XIANG ZH Q, TANG J, et al. RGBT tracking via noise-robust cross-modal ranking [J]. IEEE Transactions on Neural Networks and Learning Systems, 2022,99:1-13.

[3] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[J]. 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014:580-587.

[4] GIRSHICK R. Fast R-CNN[J]. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 1440-1448.

[5] REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017,39(6): 1137-1149.

[6] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020,42(2): 386-397.

[7] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[J]. Computer Vision-ECCV 2016, 2016: 21-37.

[8] WANG R J, LI X, LING C X. Pelee: A real-time object detection system on mobile devices[J]. ArXiv preprint arXiv:1804.06882,2018.

[9] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J]. ArXiv preprint arXiv: 1804.02767, 2018.

[10] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017:6517-6525.

[11] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[J]. 29th IEEE Conference on Computer Vision and Pattern Recognition, 2016:779-788.

[12] 许光宇,陈浩宇,张杰. 多路径生成对抗网络的红外与可见光图像融合[J]. 国外电子测量技术,2024,43(3):18-27.
XU G Y, CHEN H Y, ZHANG J. Infrared and visible image fusion based on multi-path generative adversarial network[J]. Foreign Electronic Measurement Technology, 2024,43(3):18-27.

[13] 黄月平,李小锋,卢瑞涛,等. 基于自适应标签和稀疏学习相关滤波的红外目标跟踪算法研究[J]. 仪器仪表学报,2022,43(12):199-208.
HUANG Y P, LI X F, LU R T, et al. Research on infrared target tracking algorithm based on adaptive label and sparse learning correlation filter[J]. Chinese Journal of Scientific Instrument, 2022, 43(12): 199-208.

[14] DENG Q, TIAN W, HUANG Y Y, et al. Pedestrian detection by fusion of RGB and infrared images in low-light environment[J]. 2021 IEEE 24th International Conference on Information Fusion, 2021:1-8.

[15] WU D, CAO L H, ZHOU P J, et al. Infrared small-target detection based on radiation characteristics with a multimodal feature fusion network[J]. Remote Sensing, 2022, 14(15): 3570.

[16] ZHAO Z X, BAI H W, ZHANG J SH, et al. CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion[J]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:5906-5916.

[17] 解宇敏, 张浪文, 余孝源,等. 可见光-红外特征交互与融合的 YOLOv5 目标检测算法[J]. 控制理论与应用, 2024, 41(5): 914-922.
XIE Y M, ZHANG L W, YU X Y, et al. YOLOv5 object detection algorithm with interactive and fusion of visible light-infrared characteristics[J]. Control Theory & Applications, 2024, 41(5): 914-922.

[18] SUN Y M, CAO B, ZHU P F, et al. DetFusion: A detection-driven infrared and visible image fusion network[J]. Association for Computing Machinery, 2022: 4003-4011.

[19] HE K M, ZHANG X Y, REN SH Q, et al. Deep residual learning for image recognition[J]. 29th IEEE Conference on Computer Vision and Pattern Recognition, 2016:770-778.

[20] 郑红,郑晨,闫秀生,等. 基于剪切波变换的可见光与红外图像融合算法[J]. 仪器仪表学报,2012,33(7):1613-1619.
ZHENG H, ZHENG CH, YAN X SH, et al. Optical and infrared image fusion algorithm based on shearlet transform[J]. Chinese Journal of Scientific Instrument, 2012,33(7):1613-1619.

[21] 郝洪涛,王凯,张炳建,等. 多尺度特征自适应融合的气动控制阀故障诊断[J]. 仪器仪表学报,2023,44(10):167-178.
HAO H T, WANG K, ZHANG B J, et al. Multi-scale characteristics of pneumatic control valve fault diagnosis of the adaptive fusion[J]. Chinese Journal of Scientific Instrument, 2023, 44 (10): 167-178.

[22] 胡久松,刘张驰,余谦,等. 融入 GhostNet 和 CBAM 的 YOLOv8 烟雾识别算法[J]. 电子测量与仪器学报,2024,38(8):201-207.
HU J S, LIU ZH CH, YU Q, et al. Smoke recognition algorithm of YOLOv8 integrated with ghostnet and CBAM[J]. Journal of Electronic Measurement and

Instrumentation,2024,38(8):201-207.

[23] DAI Y M, GIESEKE F, OEHMCKE S, et al. Attentional feature fusion[J]. 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020: 3559-3568.

[24] 李梅,张旭东,孙锐,等. 结合深度线索和几何结构的稀疏光场密集重建[J]. 电子测量与仪器学报,2023,37(3):1-10.
LI M, ZHANG X D, SUN R, et al. Sparse light field dense reconstruction based on depth cues and geometric structure[J]. Journal of Electronic Measurement and Instrumentation,2023,37(3):1-10.

[25] LIU J Y, FAN X, HUANG ZH B, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detectio[J]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022: 5792-801.

[26] XU H, MA J Y, JIANG J J, et al. U2Fusion: A unified unsupervised image fusion network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022,44(1): 502-518.

作者简介

李学钊,2023 年于哈尔滨工程大学获得学士学位,现为哈尔滨工程大学硕士生,主要研究方向为自动驾驶感知、多源信息融合和零样本学习等。
E-mail:1018063997@qq.com

Li Xuezhao received his B. Sc. degree in 2023 from Harbin Engineering University. Now he is a master student at Harbin Engineering University. His main research interests include autonomous driving perception, multi-source information fusion, and zero-shot learning, etc.

王伟,2001 年于哈尔滨工程大学获得学士学位,2004 年于哈尔滨工程大学获得硕士学位,2006 年于哈尔滨工程大学获得博士学位,现为哈尔滨工程大学教授,主要研究方向为智能导航与探测技术、信息融合理论和环境感知与智能决策等。
E-mail:wangwei407@hrbeu.edu.cn

Wang Wei received his B. Sc. degree in 2001 from Harbin Engineering University, received his M. Sc. degree in 2004 from Harbin Engineering University, received his Ph. D. degree in 2006 from Harbin Engineering University. Now he is a professor at Harbin Engineering University. His research interests include intelligent navigation and detection technology, information fusion theory and environmental perception and intelligent decision-making, etc.

薛冰(通信作者),2003 年于哈尔滨工程大学获得学士学位,2006 年于哈尔滨工程大学获得硕士学位,2013 年于哈尔滨工程大学获得博士学位,现为哈尔滨工程大学副教授,主要研究方向为无线电导航、惯性导航及相关多源信息融合技术等。
E-mail:xuebing@hrbeu.edu.cn

Xue Bing(Corresponding author) received his B. Sc. degree in 2003 from Harbin Engineering University, received his M. Sc. degree in 2006 from Harbin Engineering University, received his Ph. D. degree in 2013 from Harbin Engineering University. Now he is an associate professor in Harbin Engineering University. His main research interests include radio navigation, inertial navigation and related multi-source information fusion technology, etc.