可见光-红外图像融合的目标检测综述

朱自文[1+],宋晓鸥[2],崔 巍[1],岂峰利[2]
1.武警工程大学 研究生大队,西安 710086
2.武警工程大学 信息工程学院,西安 710086
+ 通信作者 E-mail:1966019840@qq.com

摘 要:随着人工智能技术的快速发展,目标检测与识别的地位日益凸显。基于深度学习的可见光-红外图像融合的目标检测技术具有强大的特征提取和泛化能力,能够有效提取和融合可见光与红外图像特征。对基于双模态图像融合检测的发展现状进行概述,并在基于深度学习的目标检测基础上分析双模态图像融合检测的优势,对比介绍常用的数据集和主要的技术难题。对基于不同阶段融合的目标检测算法进行总结分析,指出特征级融合检测的优势与主导地位;重点对基于不同基础模型的融合检测算法进行分析和总结,探讨了Transformer在双模态融合检测领域的优势和主导地位,以及Mamba在未来研究中的巨大潜力。根据当前可见光-红外图像融合的目标检测研究现状,对未来以实际的开发应用为导向进行了展望。

关键词:深度学习;目标检测;双模态融合检测;图像融合
文献标志码:A **中图分类号:**TP181 doi:10.3778/j.issn.1002-8331.2501-0206

Review of Visible and Infrared Image Fusion for Intelligent Object Detection

ZHU Ziwen[1+], SONG Xiao'ou[2], CUI Wei[1], QI Fengli[2]
1.Graduate Student Brigade, Engineering University of PAP, Xi'an 710086, China
2.School of Information Engineering, Engineering University of PAP, Xi'an 710086, China

Abstract: With the rapid development of artificial intelligence, object detection and recognition have become increasingly important. Deep learning-based object detection techniques that fuse visible and infrared images demonstrate robust feature extraction and generalization capabilities, effectively integrating features from both modalities. This paper first reviews the current state of dual-modal image fusion for object detection. It then analyzes the advantages of dual-modal fusion within deep learning-based detection and compares commonly used datasets and key technical challenges. Next, the paper summarizes object detection algorithms based on different fusion stages, emphasizing the benefits and dominance of feature-level fusion. It further analyzes fusion detection algorithms based on different base models, highlighting the advantages and dominant role of the Transformer and the potential of Mamba for future research. Finally, the paper provides a forward-looking perspective on future research oriented towards practical applications.

Key words: deep learning; object detection; dual-modal fusion detection; image fusion

目标检测是计算机视觉领域的核心任务之一,同时也是实例分割[1-2]、目标跟踪[3-4]等视觉任务的基础,目前已广泛应用于无人驾驶[5-6]、智能监控[7]、工业质检[8-10]等各个领域。作为计算机视觉的下游任务,目标检测不仅要求计算机能够识别出图像或视频中的感兴趣目标,更要实现对目标的精准定位和分类[11]。随着硬件计算能力的不断突破,深度学习在计算机应用的各个领域取得显著进展,尤其在视觉领域,基于深度学习的目标检测算法现已成为该领域发展的主流技术。

早期的目标检测技术主要依赖可见光或红外等单一模态图像作为输入,然而单一模态的图像具有信息局限性,无法为计算机提供完整的特征表达,因此通过多模态之间的信息互补可以弥补单模态信息不足的缺点。当前已有学者研究多模态图像融合的目标检测,尝

基金项目:国家自然科学基金(61801516)。
作者简介:朱自文(1997—),男,硕士研究生,CCF学生会员,研究方向为深度学习、目标检测;宋晓鸥(1983—),女,博士,副教授,研究方向为卫星导航信号处理、认知无线电;崔巍(1981—),男,硕士研究生,研究方向为军事智能与态势感知;岂峰利(1991—),男,硕士,助教,研究方向为信号处理。
收稿日期:2025-01-13 **修回日期:**2025-03-10 **文章编号:**1002-8331(2025)17-0017-16

试在确保检测精度的同时轻量化网络以实现检测的准确性和实时性[12-17]。

图1所示为LLVIP数据集[18]的红外-可见光图像对，由图可知图像对之间具有明显的互补关系。可见光图像通过捕捉可见光谱范围内的光波生成图像，与人体自然感官所接收到的信息类似。可见光图像细节纹理清晰，色彩对比度强，但抗干扰性较差，在雾雨天气和夜间拍摄的图像所能表达的信息量有限，往往会损失大量信息[19]；红外图像是通过捕捉物体自身辐射出的红外光波生成图像，受环境变化的影响较小，即使在恶劣条件下也能有效捕获目标的位置和轮廓信息，缺点是细节纹理信息丢失严重，空间分辨率低[20]。因此，通过融合可见光与红外图像的互补信息，可以有效克服复杂或恶劣环境条件下单一模态图像检测精度低的缺点，从而实现鲁棒性更强的目标检测[21]。

图1　可见光-红外图像对

Fig.1　Visible-infrared image pair

当前关于可见光-红外图像融合的目标检测综述较少，大多数集中在可见光-红外图像的融合技术综述。本文旨在系统分析当前基于深度学习的双模态融合目标检测算法的研究进展，主要内容包括：

(1)对可见光-红外图像融合的目标检测进行概述，涵盖目标检测、基于图像融合的目标检测的研究现状与分类，以及常用数据集和面临的技术难题；

(2)从不同融合阶段和不同基础模型两个角度，对基于深度学习的融合检测算法进行分类和详细阐述；

(3)基于当前可见光-红外图像融合的目标检测研究现状，对未来融合检测的发展趋势进行展望。

与以往单纯从融合阶段或时间线角度出发的综述不同，本文从融合阶段和基础模型两个角度展开，重点是对基于不同基础模型的融合检测网络进行分析和总结，旨在为未来在该领域取得创新性进展提供指导。

1　双模态图像融合检测研究现状

1.1　基于深度学习的目标检测

早期的目标检测(2014年以前)主要是基于传统的图像处理技术，依赖手工设计的特征提取，效率低下且检测精度有限。2014年，Girshick等人[22]首次提出R-CNN(region-convolutional neural network)，基于深度网络的目标检测算法，显著提高了目标检测的精度，与传统目标检测方法相比具有突出优势。

基于深度学习的目标检测网络根据检测步骤可分为单阶段目标检测和两阶段目标检测[23]，对比分析如表1所示。两阶段检测包括区域提议和检测两个步骤。首先，模型在图像中识别出可能包含目标的区域，称为区域提议或候选区域；然后，在这些区域内进行分类和边界框回归，代表网络有Fast-RCNN[24]、Faster-RCNN[25]等。单阶段目标检测将检测任务作为一个整体，直接对目标进行定位和分类，无须区域提议，代表网络为YOLO(you look only once)[26-29]系列。2020年，Carion等人[30]提出一种基于Transformer[31]的新型目标检测架构DETR，实现了端到端的目标检测，并在精度和实时性方面取得了显著优势。DETR的出现打破了两阶段和单阶段目标检测的垄断地位，首次将Transformer架构应用于目标检测领域。

Transformer最初在自然语言处理(NLP)领域得到应用，随后被引入图像处理领域。Transformer出色的自注意力机制能够从全局捕获特征之间的长距离依赖关系，有效弥补了卷积神经网络(convolutional neural network，CNN)感受野不足的缺陷。在目标检测等视觉处理任务中，Transformer展现了巨大潜力，吸引了众多学者开展相关研究[32-36]，并逐渐表现出在该领域的主导地位。

早期基于深度学习的目标检测主要集中在可见光、红外等单模态目标检测，然而随着应用场景复杂性的增加和技术要求的提高，单模态目标检测已很难满足不断提高的精度需求和环境适应性需求。多模态图像融合的目标检测由于融合了多源信息，展现出比单模态目标检测更高的检测性能和鲁棒性，因此吸引了大量学者的关注和研究。当前主流的多模态融合检测算法主要包括可见光-红外[37-38]、可见光-雷达[39]等，其中基于可见光-红外图像融合的目标检测在检测性能和环境适应能力方面表现更为出色。

1.2　基于图像融合的目标检测

图像融合技术[40]旨在充分利用不同模态图像的互

表1　单阶段目标检测与两阶段目标检测对比

Table 1　Comparison between one-stage and two-stage object detection

检测算法	区域提议	检测速度	检测精度	训练时间	参数量	代表算法
单阶段目标检测	否	快	低	短	少	OverFeat、SSD YOLO、DETR
两阶段目标检测	是	慢	高	长	多	R-CNN、SPP-Net Fast-RCNN、Faster-RCNN

补信息,通过特定的方法或手段提升融合图像的质量[41-44]。图像融合技术可以分为两类:一类是传统的图像融合技术,包括多尺度变换、稀疏表示、压缩感知、子空间表示以及显著性图像融合等方法;另一类是基于深度学习的图像融合技术[45],包括卷积神经网络、自编码器(Autoencoder)、Transformer等方法[12]。选择性状态空间模型(Mamba)的提出进一步拓宽了图像融合的途径[46]。

利用可见光-红外图像融合实现目标检测,主要是在传统基于深度学习的目标检测框架中嵌入图像融合模块,实现不同模态图像之间信息的充分交互与融合,从而提高目标检测的精度和鲁棒性[11]。目前,基于多模态图像融合的方法已成为提升目标检测性能的重要途径之一。基于深度学习的双模态图像融合目标检测按照融合与检测的关系可以分为两类:

(1)生成融合图像的检测,即首先利用图像融合技术生成高质量的融合图像,然后将其应用于目标检测等下游任务。然而,该方法实现的往往是图像特征的融合,关注是如何提升融合图像的质量,因此难以融合提取出适合计算机处理的高级语义特征[47]。为了解决这一问题,一些学者尝试将下游任务的应用需求融入到图像融合过程并指导图像融合,如Shopovska等人[48]是最早开展图像融合下游应用的研究之一,通过在损失函数中采用辅助行人检测误差的方法来帮助定义人类外观的相关特征;Liu等人[49]将图像融合和目标检测公式化为双层优化公式,并提出了一种联合训练策略,同时训练融合模型和检测模型,取得了不错的融合效果和检测效果。

(2)不生成融合图像的检测,即考虑目标检测的精度而不是融合图像的质量,在处理过程中不会生成融合图像,而是直接输出检测结果。如Hou等人[50]在YOLOv7的基础上提出一种双支路目标检测网络,利用可见光多尺度特征融合模块与红外多尺度特征融合模块,提取多个尺度的可见光与红外特征。此外,提出一种新的跨模态特征融合模块,作用于模态内从而减少模态间的冗余信息,达到提高目标检测精度的目的。该网络无论是在特征提取还是融合阶段,皆是以检测损失来训练网络,与是否生成高质量的融合图像无关。

当前主流的多模态融合检测皆是基于融合特征,而不依赖生成融合图像,这类算法不生成融合图像,而是更加注重是否能够提取出适合计算机处理的高级语义特征,实现双模态特征更好地交互和融合。大量实验证明,不生成融合图像的目标检测网络所表现出的性能要普遍好于生成融合图像的目标检测网络。

1.3 常用的可见光-红外图像数据集

基于深度学习的目标检测网络需要大量的数据进行训练,训练结果的好坏将直接影响检测的精度。利用可见光-红外图像融合的目标检测更加依赖合适的数据集,下面列举了一些常用的可见光-红外图像融合的目标检测数据集:

KAIST数据集由Hwang等人[51]于2015年创建,是广泛应用于多光谱行人检测的基准数据集之一,包含大量日间和夜间的街道、校园等场景的图像。该数据集提供大量对齐的可见光和红外图像对,分辨率分别为640×480和320×256,训练时需统一调整为640×480。原始的KAIST数据集包含95 328个被手动注释标签的配对图像,103 128个密集注释和1 182个唯一的行人注释。数据集分为50 172个图像对的训练集和45 156个图像对的测试集。需要注意的是,该数据集包含大量重复冗余的图像,使用前根据需要进行一定的清理。

LLVIP数据集由Jia等人[18]于2021年创建,是一个专门用于低光照场景下的可见光-红外配对数据集。该数据集包含16 836对配对图像,在时间与空间上严格对齐,图像绝大部分拍摄于夜间,适用于夜间或低光照条件下网络的训练和测试。该数据集包含大量行人标签信息,适合行人检测的网络模型进行训练和测试,可见光图像的分辨率达到1 920×1 080,红外图像分辨率达到1 280×720,属于高质量的可见光红外图像配对数据集。

FLIR数据集[52]由FLIR公司于2018年创建,广泛应用于可见光-红外图像融合的目标检测网络的训练和测试。数据集包含10 288个未对齐的图像对,其中8 862对图像作为训练集,1 366对图像作为测试集。可见光图像分辨率达到1 280×1 024,红外图像分辨率达到640×512,数据集的检测对象包括15个类别,囊括人、自行车、路牌、狗等检测对象,涵盖白天和夜晚的街道和公路等场景。该数据集仅对红外图像进行了注释,且可见光图像和红外图像属于未对齐状态,可用于基于未对齐图像的融检测网络训练和测试。

M^3FD数据集是Liu等人[49]于2022年创建,由校准的可见光和红外传感器同步成像系统收集而成,囊括了不同光照、季节和天气条件下的多种环境,具有丰富的像素级别的图像变化。该数据集包含严格对齐的可见光红外图像4 200对和23 635个注释过的对象,涉及街道、公路、校园、森林等各种场景,其中夜间场景1 671个,检测对象数达33 603个,分辨率达到1 024×768。

除上述数据集外,其他广泛应用的数据集如VEDAI数据集[53]、DroneVehicle数据集[54]、DVTOD数据集[55]等,各个数据集的基本情况已列于表2中,研究人员可根据所设计的目标检测网络灵活选用合适的数据集。

1.4 双模态图像融合检测技术难点

近年来,基于可见光-红外图像融合的目标检测技术发展迅速,取得了诸多重要进展,不断突破传统方法的局限。然而,该领域仍面临诸多亟待解决的技术难题,主要包括以下几个方面:

(1)照明条件的影响。光照条件恶劣的低光环境

表2　可见光-红外图像数据集

Table 2　Visible-infrared image datasets

数据集	年份	分辨率	图像对	对齐	分类	适用场景
KAIST[51]	2015年	可见光640×480 红外320×256	95 328	×	3	无人驾驶
LLVIP[18]	2021年	可见光1 920×1 080 红外1 280×720	16 836	√	1	夜间监控
FLIR[52]	2018年	可见光1 280×1 024 红外640×512	10 288	×	4	无人驾驶
M^3FD[49]	2022年	1 024×768	4 200	√	6	监控、无人驾驶
VEDAI[53]	2016年	1 024×1 024	6 000	√	9	无人机
DroneVehicle[54]	2022年	840×712	28 439	√	5	无人机
DVTOD[55]	2024年	1 090×1 080	2 179	×	3	无人机

下,可见光图像存在严重的信息缺失,导致在图像融合过程中引入大量干扰信息[56]。因此,如何有效平衡红外和可见光图像在不同照明条件下的融合权重,成为多模态融合目标检测的一个关键难点。针对照明条件对融合检测的影响,Yan等人[57]提出一种自适应的光照感知权重生成模块,通过感知各种光照条件,实现对可见光和红外图像最终检测置信度的自适应加权,提高融合检测的鲁棒性。

(2)数据对齐与配准。当前提出的双模态融合检测网络大都是利用图像配准对齐的数据集,图像之间在像素层面有很强的对应关系,然而在实际应用中采集的不同模态图像很难实现严格的对齐。不同模态图像在获取之后,通常处于弱对齐状态,即不同模态图像会存在一定的位置偏移,给网络的训练带来困难,例如使用共享的边界框来匹配两种模态的对象将会变得更加复杂[58]。

(3)特征融合的挑战。实现特征级图像融合时,需充分考虑到不同模态图像的互补信息和特征差异性,针对该问题已有许多融合策略提出,但是如何充分利用两种模态的互补信息来提高目标检测的性能,仍旧需要不断地改进和摸索,目标检测的性能仍有进一步提升的空间。

(4)深度学习模型的压缩与实时性。目标检测技术无论是民用还是军用,对目标检测的实时性都提出很高的要求。随着目标检测技术的快速发展,网络的深度和宽度在逐渐增大,与此同时使用双模态图像作为输入也会大大增加模型的复杂度和参数量[59],这对网络的实时性提出巨大的挑战。如何在引入双模态图像融合的同时保持目标检测的实时性与网络结构的轻量化,已成为摆在研究人员面前的一个突出难题。

2　基于深度学习的双模态融合检测网络

2.1　基于不同融合阶段的融合检测网络

基于可见光-红外图像融合的目标检测根据融合阶段的不同,可以分为像素级融合、特征级融合和决策级融合[60],不同阶段的融合策略拥有各自的优势和局限性,适用于不同的应用场景。

像素级融合检测涉及两种模态图像像素级别的信息交互,一般在网络的初始阶段,目的是提取信息更为丰富的细节特征。像素级融合的最大优势是能够最大程度的保留和利用原始图像的信息,由于需要逐像素处理,导致融合速度较慢,且无法提取适合网络学习的高级语义特征[45]。如Wagner等人[61]首次尝试了基于深度学习的可见光-红外图像融合目标检测,采取直接在通道维度拼接源图像的方法,送入特征提取网络进行特征提取。然而,实验证明该方法效果一般,原因是缺乏高级语义特征来训练网络。尽管像素级融合的目标检测网络具有一定的局限性,但是仍有学者尝试利用像素级实现图像特征的融合[62],像素级融合的目标检测网络框架如图2所示。

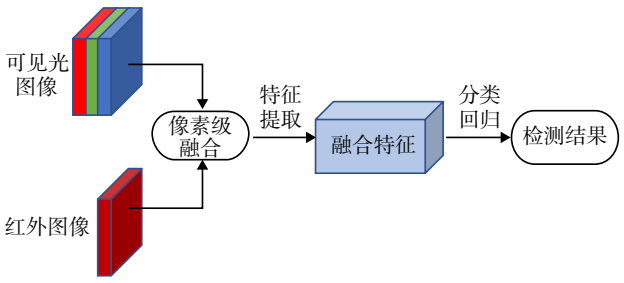


图2　像素级融合目标检测框架

Fig.2　Pixel-level fusion object detection framework

特征级融合属于较高层级的融合,主要发生在特征提取阶段,网络框架如图3所示。基于深度学习的检测网络具有出色的特征提取能力,能够从原始图像中提取出高级语义特征,包含图像的纹理、边缘等细节信息[60],因此在特征级别进行特征的融合,能够充分交互两种模态之间的高级语义特征。特征级融合的检测网络能够达到很高的检测精度,但关键就是是否能够设计出合适的特征融合方法,如Shen等人[63]在特征提取阶段的中期通过引入注意力机制,有效捕捉两种模态的互补特征,实现双模态高级语义特征的多次融合;Xie等人[64]基于YOLOv5框架设计的图像融合检测网络,将C3模块替代为特征交互模块,实现特征提取中期双模态高级语义

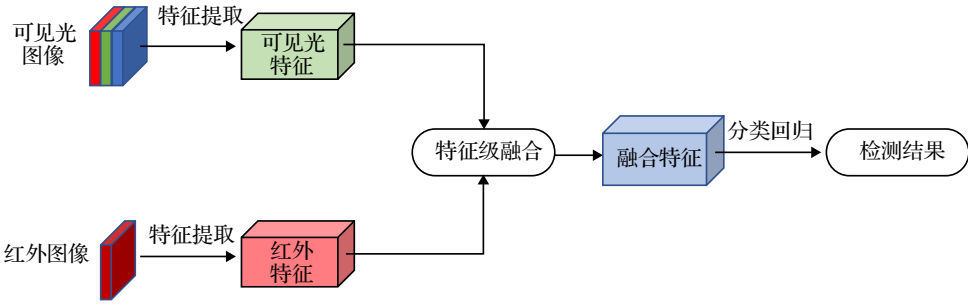


图3 特征级融合目标检测框架

Fig.3 Feature-level fusion object detection framework

特征之间的多次交互。

决策级融合检测是一种在决策阶段进行图像融合的目标检测网络,包括加权图像融合[65-66]以及决策组合[67]等,网络模型如图4所示。单模态图像首先在网络中进行特征的提取和分类回归,然后将结果组合输出实现对目标的检测,通常包括置信度得分以及目标框的回归值等[60]。基于决策级融合的目标检测网络能够在决策阶段充分考虑单模态检测的最佳结果,从而克服单一模态判决的局限性,提高检测网络的准确性和鲁棒性。例如,Li等人[66]考虑到不同模态图像在不同光照条件下有用信息贡献度不同,采取特征级与决策级融合相结合的方法,利用置信度检测分支来为最终决策分配置信度权重,实现检测结果的加权融合。然而该网络虽然考虑了决策阶段不同模态信息的贡献程度,但是却忽略了在特征融合阶段同样存在贡献程度的不同,导致不同模态信息并没有得到充分利用。

像素级融合、特征级融合和决策级融合都是为了实现两种模态之间信息的交互,充分利用双模态融合检测的双源优势实现目标检测。像素级融合可以最大程度保留源图像的信息,但却很难提取出适于计算机处理的高级语义融合特征;特征级融合能够实现高级语义级别的信息交互,但是却存在细节信息丢失、算法设计复杂和可解释性差等缺点;决策级融合灵活性高,模态检测具有独立性,但是却存在双模态图像的共享和差异信息无法充分融合的缺点。不同融合阶段的检测网络优缺点对比如表3所示。由于融合效果的好坏将直接导致检测精度的高低,因此研究与创新大多发生在特征级融合,但是面对复杂的应用环境,研究者也可灵活选择融合阶段或采取多个阶段融合等。

2.2 基于不同基础模型的融合检测网络

2.2.1 CNN

随着深度学习的兴起,卷积神经网络(CNN)在目标检测领域不断取得新的突破[22,24-25]。CNN具有强大的局部特征提取能力,能够捕获图像的细节和纹理特征,是多模态融合检测算法的基础和核心。此外,CNN还具有参数量低、运算效率高等优点。

2015年,Hwang等人[51]提出KAIST数据集,推动了可见光-红外图像融合的目标检测研究。2016年,Wagner等人[61]提出一种用于行人检测的可见光-红外图像融合的深度学习网络,该网络首次尝试了基于像素级和特征级融合的目标检测,网络结构如图5所示。实验表明,像素级融合难以学习到图像的高级抽象特征,检测性能明显低于特征级融合。同年,Liu等人[71]基于Faster R-CNN提出了四种融合结构,将两个分支的卷积神经网

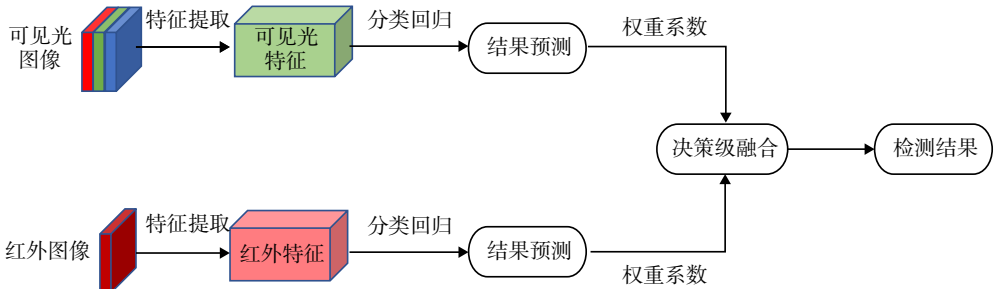


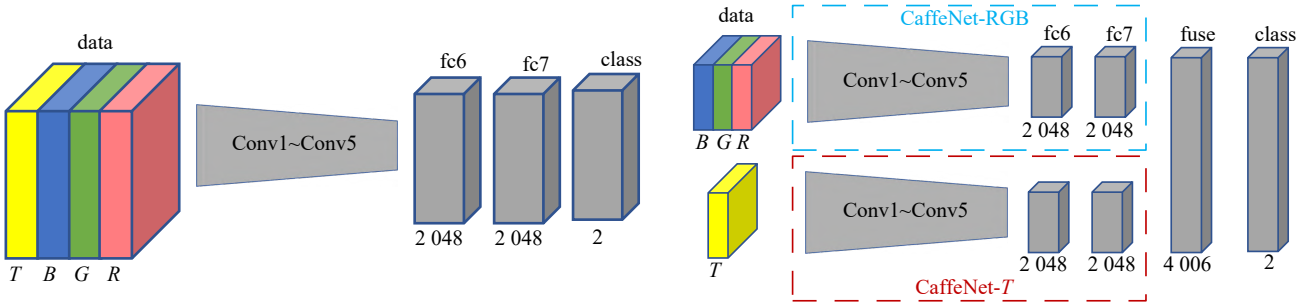
图4 决策级融合目标检测框架

Fig.4 Decision-level fusion object detection framework

表3 不同融合阶段对比

Table 3 Comparison of different fusion stages

融合阶段	优点	缺点	相关文献
像素级	像素级别信息交互,保留源图像融合信息	缺乏高级语义特征,运算量大等	文献[62]
特征级	高级语义特征级别信息交互,注意力机制	丢失像素信息,可解释性差等	文献[38,60,63,68-69]
决策级	模态独立性,灵活性高和易于实现	无法充分利用共享和差异信息等	文献[63,67,70]



（a）像素级融合　　（b）特征级融合

图5　可见光-红外双模图像融合检测网络

Fig.5　Visible-infrared dual-modal image fusion detection network

络集成在卷积层、全连接层和决策层等不同阶段，分别对应低级、中级、高级和置信度信息的融合。实验表明，中期融合可以更好地实现不同模态信息的融合，为最终检测提供充分的特征表达。然而，该网络并未考虑到不同模态图像的共享与差异信息，仅是简单的特征融合，导致其无法充分发挥双模态融合检测的双源优势。

2020年，Zhang等人[72]指出，即使可见光与红外图像在空间上对齐，提取后的特征也会存在不一致，因此他们提出一种基于循环融合和特征优化的检测网络，旨在增强不同模态之间的特征一致性，网络结构如图6所示。研究表明，特征融合的效果与循环融合的次数密切相关，当次数为3时效果最好，大于3时模态间过多的一致性会导致融合失去意义。2024年，Kang等人[68]提出的全局-局部特征融合网络（global-local feature fusion，GLFNet）同样基于循环融合结构。GLF特征融合模块能够自适应的从单模态特征中提取显著信息，具有即插即用、扩展性强的优点。在DroneVehicle数据集上，GLFNet的平均精度均值（mAP）达到70.7%，较UA-CMDet[54]提升了7.4%。循环多次融合结构能够在多个尺度提取和融合特征，实现对单模态特征的增强，近年来已涌现大量类似网络[34,57,64,73-74]。

注意力机制首先在自然语言处理领域取得成功，后来被引入到计算机视觉领域，包括空间注意力机制（spatial attention mechanism，SAM）[75]、通道注意力机制（channel attention mechanism，CAM）[76]、混合注意力机制[77]和自注意力机制[31]等。2023年，Yang等人[78]提出一种级联信息增强和跨模态特征融合网络，该网络利用CAM和SAM对拼接后的融合特征进行进一步处理，并通过与单模态特征相乘，实现在通道和空间维度的信息增强。此外，设计了一种跨模态注意力特征融合模块，通过不同模态特征在通道维度的互补增强，实现对融合特征的进一步优化，显著提高了对感兴趣目标的关注度。在KAIST多光谱数据集上，该检测网络将漏检率分别降到10.71%（全天候）、13.09%（白天）和8.45%（夜间）。消融实验证明，级联信息增强模块和跨模态特征融合模块分别可以降低2.63%和2.16%的漏检率，这进一步证明了注意力机制在多模态融合检测中的有效性。

2024年，Fu等人[38]为了充分利用不同通道之间的远程相关性，设计了一种特征增强的远程注意力融合网络（long-range attention fusion network，LRAF-Net）。提出了一种非对称互补掩膜的数据增强方法，通过随机遮挡局部信息，迫使网络更加关注不同模态的互补信息，从而减少对单一模态的依赖。此外，LRAF-Net设计了一种通道和空间注意力模块，从多个尺度对特征进行融合，利用通道和空间维度上的互补信息增强融合特征。同时基于Swin-Transformer[79]设计了一种远程依赖融合

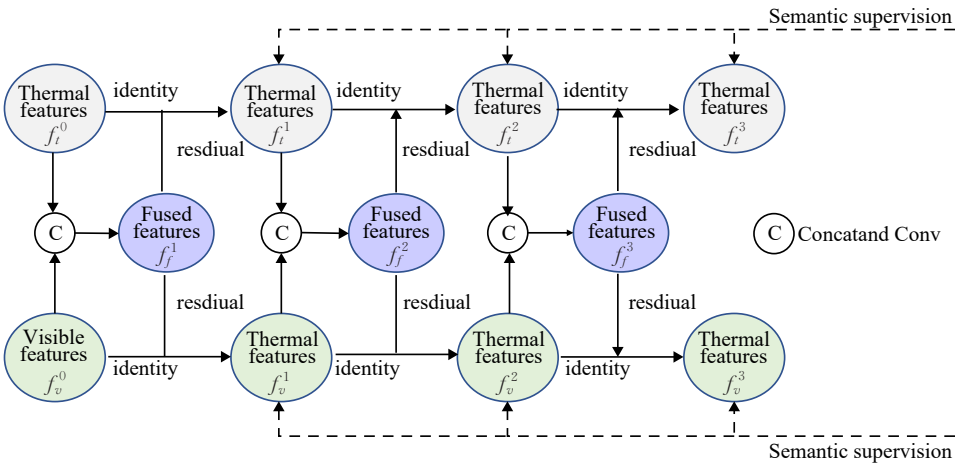


图6　循环多次融合结构

Fig.6　Recurrent multi-fusion structure

模块,利用位置编码的远程相关性实现互补信息的融合,解决了传统方法仅依赖局部信息融合导致的性能不足问题。LRAF-Net在VEDAI、FLIR和LLVIP数据集上的mAP分别可以达到59.1%、42.8%和66.3%,检测精度显著提升。然而,该网络的高计算成本和对数据增强策略的过度依赖,限制了其在实际中的应用推广。注意力机制在可见光-红外图像融合的目标检测中应用广泛[38,68-69,80-82],并且至今仍旧有较高的竞争力。

光照条件的变化会严重影响基于可见光-红外图像融合的目标检测性能。2019年,Li等人[56]研究了光照条件变化对可见光-红外图像融合目标检测性能的影响,实验发现光照条件良好的环境下基于可见光图像的检测性能要优于红外图像;而在夜间,基于红外图像的检测性能要优于可见光图像,甚至超过融合检测的性能。由此,Li等人[56]认为,可见光图像在光照条件不良时对检测做出的贡献可能小于其引入的干扰。为解决这一问题,提出一种基于照明感知的IAF(illumination-aware faster)R-CNN网络,该网络由多光谱检测网络Faster R-CNN、光照分支估计模块和门控融合模块构成,能够实现不同光照条件下可见光-红外图像的决策级融合检测。尽管决策级融合是当前光照感知网络发展的主流[65,67],但该方法无法充分融合不同模态的特征信息,具有一定的局限性。

程清华等人[83]针对不同光照条件下图像的贡献差异,提出一种基于光照感知的红外-可见光融合目标检测方法。该方法提出一种轻量化的光照感知权重分配模块,与以往方法不同的是,该模块设置在特征提取的中期,用于调整不同模态特征在融合时的贡献程度,从而提高了融合检测算法在不同光照条件下的适应性和准确性。此外,提出了一种无参数的3D注意力模块,能够自动识别特征的通道和空间重要性,并根据模态间的相对重要性分配不同的融合权重。该检测网络在NRS和M^3FD数据集上的mAP可以达到71.2%和63.1%,显著优于当前绝大多数目标检测网络。然而,该方法仅考虑了融合权重的比例而并未考虑多模态图像的差异信息,从而限制了其精度的提升,此外,该网络在更多场景下的应用也未得到充分验证。

基于可见光-红外图像融合的目标检测大多依赖严格对齐的图像对进行训练,然而获取大量严格对齐的配对图像成本极高,训练出的网络也难以得到应用,因此出现了弱对齐的可见光-红外图像融合的目标检测网络[84-86]。弱对齐的数据对齐模式主要包括三种:像素级对齐、区域级对齐和决策级对齐[87]。

2020年,Zhou等人[88]提出了一种名为modality balance network(MBNet)的多光谱行人检测网络,旨在解决多光谱行人检测中的模态不平衡问题。提出一种光照感知特征对齐模块(IAFA),利用可见光图像来预测光照强度,从而为可见光与红外特征分配权重;在预测光照条件的基础上,该网络设计了一种模态对齐模块,为每个模态的每个像素点(x,y)预测偏移量(dx,dy),利用双线性插值的方法从四个相邻像素中获取最终的像素值$(x+dx,y+dy)$,用于调整红外特征图,最后与可见光特征图结合生成更加平衡的特征表示。在KAIST数据集上,添加IAFA模块的MBNet整体漏检率从9.36%降低到8.13%,显著提高了目标检测的精度,但是由于需要逐像素处理,导致其计算成本激增,不利于网络的实际部署。

Zhang等人[84]探讨了多光谱行人检测中的位置偏移问题,提出一种弱对齐跨模态融合的行人检测网络。提出了一种基于区域特征对齐(region feature alignment,RFA)的模块,通过固定红外特征来预测区域间的偏移并进行特征对齐,解决了不同模态间特征不匹配的问题。此外,采取服从正态分布的兴趣区域(region of interest,RoI)抖动策略来随机抖动可见光分支中的RoI,从而增强网络对不同偏移模式的鲁棒性。最后,利用置信度感知融合模块对两种模态的特征进行加权,强化有用特征并抑制无用特征。在KAIST数据集上,该检测网络的漏检率可以达到9.94%(白天)、8.38%(夜间)和9.34%(全天候),然而RFA模块和RoI抖动策略的引入,显著增加了模型的计算复杂度,严重影响网络实时性。

2024年,Tian等人[86]提出了一种跨模态提议引导的特征挖掘机制(cross-modality proposal-guided feature mining,CPFM)检测头,旨在解决因图像未对齐导致的行人信息丢失问题。该机制能够在图像未对齐情况下,有效利用两种模态之间的互补性,并在决策阶段解决未对齐图像中行人位置不一致的问题。该方法首先预测一个能够包括两种模态图像中目标位置信息的包围框,确保不丢失任何模态的目标信息。然后,利用融合特征预测跨模态提议,并通过包围框进行监督学习,最终在模态提议的指导下进行特征挖掘,用于回归每个模态中的精确边界框。为了进一步改善未对齐图像的处理,作者还设计了一种基于单应性变换的数据增强算法,使模型能够接触到更多样化的未对齐图像对,从而提高对未对齐图像处理的鲁棒性。在KAIST数据集上,该检测网络相对于其他以VGG-16和ResNet-50为骨干网络的模型,分别提高了26.8%和13%的检测性能。尽管该模型在模拟偏移的实验中表现出良好的鲁棒性,但在实际部署中,若图像未对齐程度超过模型的训练范围或者存在一定的遮挡时,检测的精度可能会出现一定程度的下降,因此,该网络仍有进一步优化的空间。

针对图像存在平移、缩放和旋转等复杂偏差的情况,Fu等人[87]提出一种快速的单阶段检测网络YOLO-Adaptor,通过引入轻量级的多模态适配器和特征对比学习损失,在特征提取阶段直接预测对齐参数和模态间

的置信度权重,从而有效解决了可见光和红外图像之间的复杂偏差(包括平移、缩放和旋转)。在KAIST数据集上,当图像对旋转不同角度时,YOLO-Adaptor的漏检率均低于6.18%,此外在FLIR和LLVIP数据集上的mAP50分别达到80.1%和96.5%。然而,在小目标较多的场景中,YOLO-Adaptor的性能提升相对有限。例如,在FLIRn和LLVIPn数据集的实验中,mAP50分别降低了8.4%和2.6%。

CNN凭借其强大的局部特征提取能力和高效的运算效率,成为多模态融合检测的核心。从早期直接的像素级和决策级融合,逐步发展到循环多次的特征级融合,再到空间、通道及跨模态等注意力机制的广泛引入,多模态融合检测的特征一致性与检测性能得到显著提升,特征互补性也得到增强。此外,针对光照条件变化和图像未对齐等技术难题,研究者们提出的光照感知网络和弱对齐融合策略,也在不断提高检测模型的适应性和鲁棒性。

2.2.2　Transformer

2020年,Carion等人[30]首次提出基于Transformer的端到端目标检测网络DETR,展现了Transformer在目标检测领域的巨大潜力。Dosovitskiy等人[89]提出Vision Transformer(ViT),首次将完整的Transformer架构应用于图像处理领域。ViT通过将图像分割为块序列并添加位置编码,利用自注意力机制实现全局信息的交互,从而将具有长距离依赖关系的特征联系起来,为图像的分类提供更好的特征表达。ViT的提出引起了广泛关注,并吸引了大量学者投入到Transformer在图像处理领域的研究[32,34-36,79,90],为多模态融合检测开辟了全新的思路。

2020年,Fang等人[91]首次将Transformer应用于可见光-红外图像融合的目标检测,提出了一种简单而高效的跨模态特征融合模块(cross-modality fusion Transformer,CFT),网络结构如图7所示。CFT不仅能够学习远程特征之间的依赖关系,整合上下文信息,而且Transformer的自注意力计算机制也可以轻松实现模态内和模态间的特征融合,无须设计复杂的融合结构也可有效捕捉可见光与红外图像之间潜在的互补信息,提高目标检测的性能。该网络基于YOLOv5框架,采取多次循环融合的方法,在FILR、LLVIP、VEDAI数据集上的mAP分别可以达到40.2%、63.6%、56.0%,相比基线方法分别提升了2.8%、1.3%和9.2%。尽管检测精度显著提升,但模型参数量和复杂度也显著增加,相比基线方法增加了132.31×10^6和34.26 GFLOPs,较高的计算开销导致其在实际部署中存在一定的局限性。

2023年,Lee等人[92]提出一种基于Transformer的交叉注意力融合的两阶段行人检测网络,利用交叉引用的思想以一种模态来指导提取另一种模态的潜在特征。其计算步骤如图8所示,计算公式如下:

$$Attention(\boldsymbol{Q}_r,\boldsymbol{K}_t,V_r)=softmax(\alpha\boldsymbol{Q}_r\boldsymbol{K}_t^{\mathrm{T}})V_r$$

$$Attention(\boldsymbol{Q}_t,\boldsymbol{K}_r,V_t)=softmax(\alpha\boldsymbol{Q}_t\boldsymbol{K}_r^{\mathrm{T}})V_t$$

交叉模态注意力模块可以鼓励网络学习到被单模态忽略而被另一模态突出的特征信息,具有比传统融合方式更强的互补信息融合能力。同年,You等人[93]提出了一种双流多尺度聚合网络(multi-scale aggregation network,MSANet),利用多尺度聚合Transformer在不同尺度上对多模态特征进行融合和增强。此外,该网络引入TNT[94]结构,提出内部Transformer和外部Transformer,负责提取具有局部视角的小尺度特征和具有全局信息

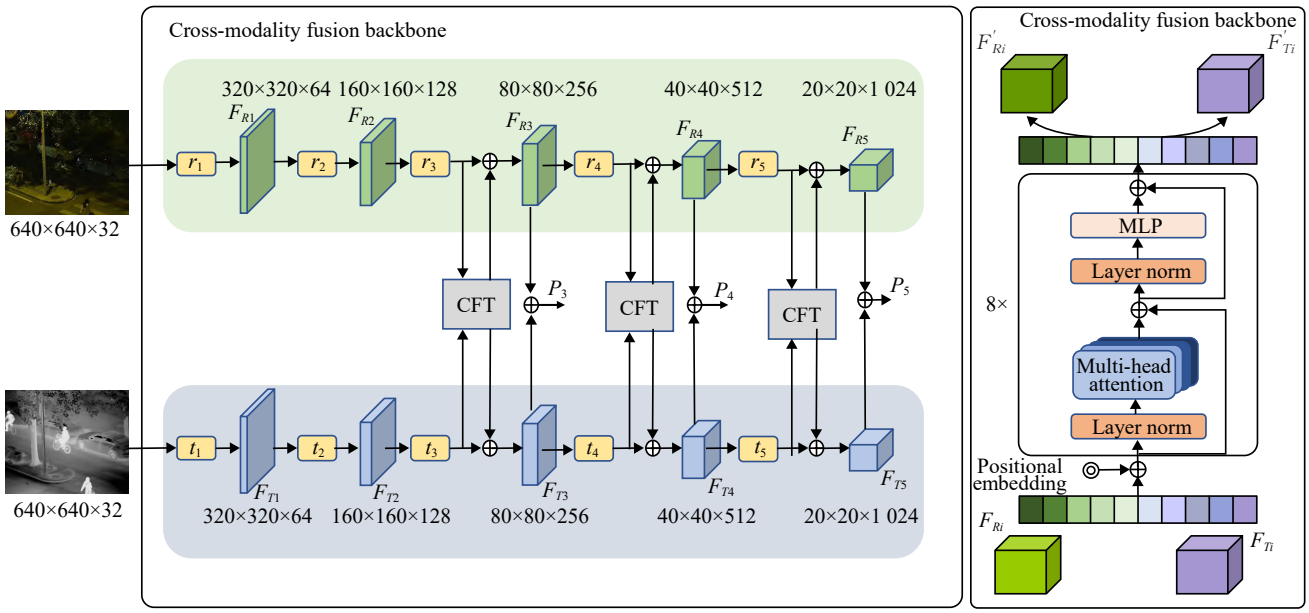


图7　基于Transformer的跨模态融合网络

Fig.7　Transformer-based cross-modal fusion network

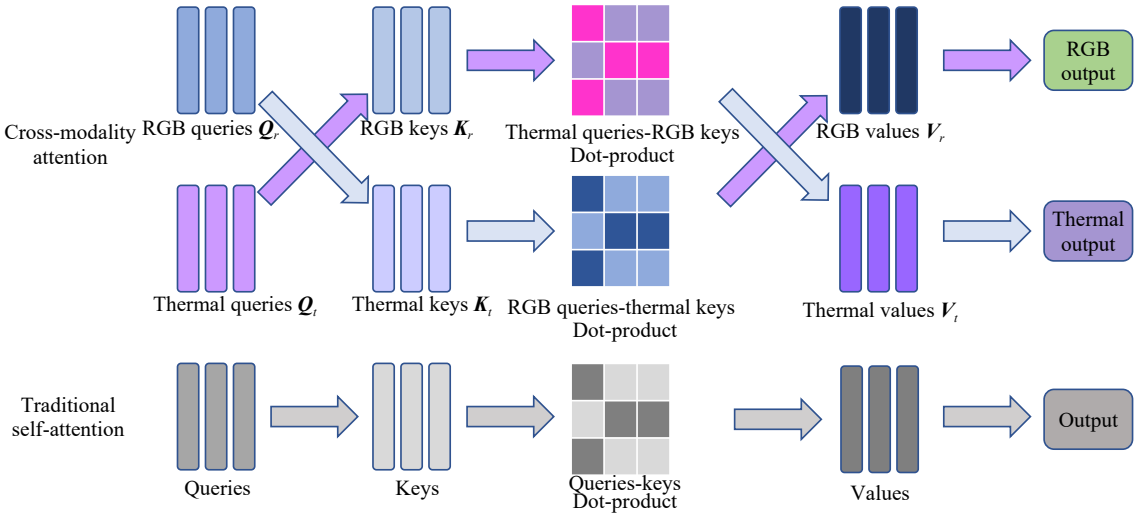


图8 交叉模态注意力模块

Fig.8 Cross-modal attention modal

的大尺度特征。通过这种设计,MSANet能够更加全面地捕捉目标的语义信息,从而提高检测精度。在FLIR数据集上,MSANet的mAP可以达到67.4%,然而TNT结构的引入显著增加了网络的参数量和计算复杂度,导致其难以在无人机、无人驾驶等资源受限的设备上进行部署。

2024年,Shen等人[63]提出一种名为ICAFusion(iterative cross-attention fusion)的多光谱目标检测网络,通过迭代交叉注意力引导特征融合,有效提高了多光谱目标检测的性能。此外,提出了一种迭代学习策略,利用共享参数在多个迭代中逐步细化特征表达,避免了多个Transformer块堆叠带来的参数和计算成本增加。ICAFusion具有较高的检测精度和推理速度,在FLIR和VEDAI数据集上的mAP50分别达到了79.20%和76.62%,推理速度可达38.46 FPS。然而,网络在背景噪声抑制、遮挡处理和输入模态质量依赖方面仍存在不足,其泛化能力有待进一步提升。

同年,Nie等人[34]提出一种CNN-Transformer级联的局部-全局跨模态特征提取模块,CNN负责提取局部细节信息,Transformer负责提取基于全局的远程关联信息。通过注意力机制动态调整不同模态特征的权重,增强了融合特征的多样性,提高了模型对跨模态信息的利用效率。此外,该网络采用并行结构分别处理可见光和红外图像,避免模态间信息的直接混合,从而保留各自模态的独特信息。在VEDAI数据集上,该网络的mAP达到79.8%,比基线方法提升了5.7%。然而,保留模态的独特信息可能导致模态间的共享信息无法充分指导特征提取,不利于融合特征的增强,同时在环境变化时网络的鲁棒性也需要进一步验证。

RT-DETR[33](real-time detection Transformer)是首个基于Transformer的实时端到端检测网络。受RT-DETR启发,Xiao等人[95]提出一种名为GM-DETR(generalized multispectral DETR)的可见光-红外图像融合目标检测网络,旨在通过高效的特征融合和两阶段训练策略,提高可见光-红外图像融合的目标检测网络的准确性和鲁棒性。同时,DETR框架的引入使得网络能够以端到端的方式实现目标检测任务,在提高检测准确性的同时保证效率。在FLIR和LLVIP数据集上,GM-DETR达到了45.8%和70.2%的mAP,且模型参数量仅为70×10^6,计算量为176 GFLOPs,推理速度达到218 FPS,满足实时性要求。尽管GM-DETR能够做到实时性和端到端检测,但在多模态特征间的差异信息处理方面还存在不足,训练难度较大,且对小目标的检测能力有待进一步提升。

Transformer的自注意力机制可以从全局捕获远距离特征之间的关系,独特的计算机制也可轻松融合不同模态的特征信息[96]。然而,计算复杂度高、硬件要求高等问题限制了其在实际应用中的表现。与此同时,纹理、细节信息的提取能力不足等缺点,也促使研究者利用CNN进行先期处理来弥补,这在一定程度上也增加了网络的复杂度。从当前基于可见光-红外图像融合的目标检测发展趋势来看,基于Transformer的融合检测网络仍然是未来极具发展潜力的研究方向。尽管Transformer的引入显著提升了检测精度,但同时也带来了网络复杂度和计算资源消耗的增加。因此,当前研究面临的主要挑战是如何在保持高检测精度的同时,有效降低网络复杂度并提升网络在复杂环境下的鲁棒性,这些问题仍需研究人员进行深入探索。

2.2.3 Mamba

Mamba是由Gu和Dao[97]于2023年提出的一种新型状态空间模型,拥有类似于RNN的推理结构,核心思想是通过选择性状态空间(selective state space)模型,在保持线性计算复杂度的同时,实现对长序列数据的高效建模。与CNN相比,Mamba具有全局建模能力,选择性扫描机制使其具有很高的输入适应性;与Transformer

相比,Mamba在处理长序列时具有线性复杂度,计算开销远小于Transformer。CNN、Transformer和Mamba模型对比如表4所示。Yu等人[98]指出,基于Mamba的深度学习网络在长序列建模方面表现出色,但在某些视觉任务(如ImageNet图像分类任务)中并不具备显著优势。图像分类的目的是实现对目标的准确分类,其过程不依赖高分辨率的图像作为输入,因此不涉及长序列建模,此时Mamba相对于Transformer在处理长序列上的优势此时也无法得到体现。然而,对于目标检测和语义分割等任务,由于通常需要高分辨率的图像作为输入,涉及长序列的处理,此时Mamba线性增长的计算复杂度优势就能得到充分体现。尽管Mamba在图像分类任务中的表现较为有限,但是在目标检测和语义分割等任务中仍展现出巨大潜力,并已受到广泛关注[99-101]。

二维视觉数据与语言数据不同,无法直接利用状态空间模型处理,因此Liu等人[100]借鉴DETR的图像分块处理方法,提出一种4路/2-D选择性扫描方法,并在此基础上构建了基于Mamba的深度网络VMamba。2-D选择性扫描的处理过程如图9所示。2-D选择性扫描首先对输入的特征图从4个方向进行扫描,生成4个序列,确保特征图中的每个元素都可以学习到来自不同方向的关联信息。生成的序列分别输入到选择性扫描状态空间模型中进行处理,捕捉远距离特征之间的相互关系,而且其线性复杂度的优势可以使其实现快速的推理过程。VMamba在目标检测任务中表现出色,性能远高于Swin-Transformer[102],展现了极高的潜力。此外,2-D选择性扫描方法的提出,也为Mamba在多模态融合检测领域的应用开辟了道路。

2024年,Wang等人[103]提出了一种名为MGMF(mask-guided Mamba fusion)的可见光-红外图像融合的目标检测网络,用于无人机平台上的车辆检测任务。该方法通过引入掩码引导的特征融合策略,有效整合了多模态特征信息。MGMF首先根据输出的预测框生成掩码图,并输入到掩码正则化约束模块,实现对无用信息的抑制和重要特征的增强。增强后的特征输入到融合模块,经初步融合后输入到2-D选择性扫描模块,利用状态空间模型捕获特征间的远程关联性,实现双模态图像的特征级融合。在LLVIP和DroneVehicle数据集上,MGMF的mAP@0.5:0.9能够达到69.8%和55.2%。然而该网络的参数量达到了122×10^6,推理速度也仅为6.1 FPS,难以满足实时性要求。

Li等人[104]提出一种在恶劣条件下进行多光谱目标检测的交叉模态融合网络(cross-modality fusion mamba with weather-removal,CFMW),通过引入天气去除扩散模型和跨模态融合Mamba模块,有效提升了在复杂天气条件下目标检测的性能。CFMW中的交叉模态融合Mamba模块基于状态空间模型,在不同尺度进行多次特征融合和强化。首先,利用四路扫描方法将图像投影为块序列,通过通道交换Mamba模块[105]对不同模态的信息进行交互,增强模态间的相关性,随后输入到状态空间模型进行相关性处理,并利用门控融合机制对输出的图像序列进行动态融合,在保持高精度的同时,显著降低了计算复杂度。在自建的数据集SWVID上,MGMF达到97.2%的mAP50、76.9%的mAP75和63.4%

表4　CNN、Transformer和Mamba算法对比

Table 4　Comparison of CNN, Transformer and Mamba algorithms

算法模型	核心机制	上下文捕获能力	计算复杂度	参数量	运算效率	优点	缺点
CNN	卷积、池化、注意力机制	局部上下文	线性复杂度	低	高	局部特征提取强,硬件资源需求低	远距离特征提取能力差
CNN+Transformer	自注意力机制编解码器	全局上下文	二次方复杂度	高	低	全局特征提取强,并行度高	计算资源消耗大
CNN+Mamba	选择性扫描、状态空间模型	全局上下文	线性复杂度	低	高	建模能力与Transformer相当,计算复杂度低	长距离依赖处理有局限

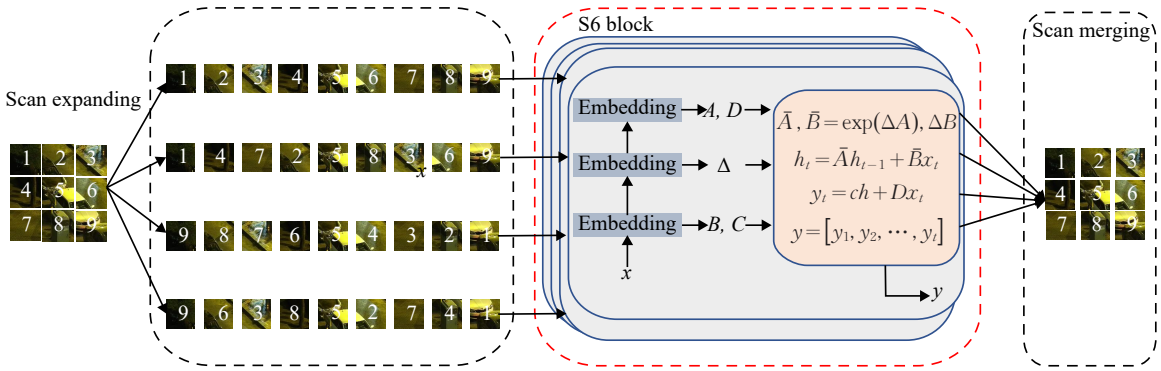


图9　2-D选择性扫描

Fig.9　2-D selective scan

的mAP,在处理高分辨率图像时,MGMF仅需10.72 GB的内存,相比CFT(21.88 GB)节省了11.16 GB。尽管CFMW在SWVID数据集上表现出色,但在其他数据集或实际应用中的泛化能力和鲁棒性仍需进一步验证。

Ren等人[106]针对无人机航空成像尺寸小、分布密集、类间分辨率低等困难,提出了一种多模态遥感探测网络RemoteDet-Mamba,旨在提高在复杂环境下的小目标检测性能。该网络通过CNN学习单模态局部特征,利用Mamba对多模态图像序列的全局特征进行融合,增强小目标的可区分性并改善类间差异。此外,RemoteDet-Mamba首次在遥感图像的目标检测中引入选择性扫描机制,利用4路扫描生成图像序列,并对不同模态同一扫描方向的图像序列采取先相加后处理的方法,实现模态间特征的深度融合。在Dronevehicle数据集上,RemoteDet-Mamba表现出色,mAP可以达到81.8%,模型参数量为71.34×10^6,检测速度达到24.0 FPS,满足实时性需求。尽管该网络在检测精度和处理速度方面表现出色,但其融合方法存在共享与差异信息利用不充分的问题,融合策略仍有进一步优化的空间,且网络在环境变化时是否还能保持较高鲁棒性有待商榷。

可见光与红外图像分别由不同的传感器捕获,因此图像之间往往会存在错位。2024年,Liu等人[107]提出了一种名为COMO(cross-mamba interaction and offset-guided fusion)的新型多光谱目标检测网络,通过引入跨模态交互和偏移引导融合策略,有效解决了多模态图像之间存在的错位问题,显著提高检测性能。COMO引入了Cross-Mamba交互模块,通过序列化的特征提取和交互,在减少计算开销的同时提高了特征的融合效果;提出一种偏移引导融合策略,通过高级特征引导低级特征的融合,实现最大化信息保留,减少错位对检测性能的影响。COMO表现出强大的泛化能力,在DroneVehicle、LLVIP和VEDAI等多个数据集上的mAP分别达到65.5%、65.2%和50.3%,而参数量和计算量仅为16.32×10^6和14.03 GFLOPs,推理速度达到227.2 FPS,表现出良好的实时性。但是,COMO高度依赖高级特征来减少错位的影响,若高级特征的提取不够充分或者丢失,则可能会限制其检测性能,且模型结构复杂,可能导致模型的训练成本增加。

除上述检测网络外,还有Dong等人[101]提出的Fusion-Mamba跨模态目标检测网络,以及Zhou等人[108]提出的DMM(disparity-guided multispectral Mamba)目标检测网络等。基于Mamba的可见光-红外图像融合目标检测网络的效率对比如表5所示。尽管Mamba模型在多模态融合检测任务中展现出了卓越的长序列建模能力和线性计算复杂度,使其在效率和性能上具有与Transformer相媲美的潜力,但当前的研究大多停留在理想条件下,对于光照变化、雨雾天气以及特殊应用场景下的适应性研究仍然不足。此外,现有研究大多是将Mamba模型直接嵌入检测网络,缺乏针对具体应用背景的定制化改进,这在一定程度上限制了Mamba在多模态融合检测中的潜力。Mamba模型尽管在多模态融合检测领域的潜力巨大,但要实现其在复杂现实背景下的实际应用,仍需在鲁棒性、适应性改进以及与现有技术融合等方面开展进一步研究。

表5 基于Mamba的检测网络效率对比

Table 5 Efficiency comparison of detection networks based on Mamba

检测网络	数据集	mAP/%	参数量/10^6	FPS
MGMF[103]	LLVIP	69.8	122.00	6.1
CFMW[104]	LLVIP	64.8	—	—
COMO[107]	LLVIP	65.2	16.32	227.2
RemoteDet-Mamba[106]	DroneVehicle	81.8	71.34	24.0
FMB[101]	LLVIP	64.3	287.60	12.8
DMM[108]	DroneVehicle	79.4	87.97	—

基于不同机制的网络模型有其独特的优势和局限性,同时每一个检测网络也存在其独特的创新方法和不足,要想完整的评价一个检测网络,必须要看其解决了什么问题以及应用于什么场景。基于不同模型的目标检测网络在机制、优势、局限性和适用场景等方面的对比见表6。

3 双模态图像融合检测发展展望

随着硬件计算能力的不断提升,基于深度学习的目标检测算法在相当长时间内仍然居于主导地位,尽管多模态图像融合检测研究已经取得显著成绩,但是仍然有巨大的发展空间,主要体现在以下几个方面:

(1)未对齐图像的处理。在实际应用中,由于传感器不同或设备老化、磨损或校准不准确,很容易导致得到的不同模态图像处于未对齐状态。尽管当前已有大量学者在进行相关的研究,但是绝大部分学者还是将注意力集中在检测精度的提升或是模型规模的轻量化改进,以及新技术在融合检测领域的探索等。图像未对齐问题的解决与否,将直接导致检测网络在实际生活中的应用,因此如何解决图像数据未对齐的问题仍然是当前融合检测领域研究的一个重要方向。

(2)图像融合与目标检测任务的结合。随着图像融合技术的不断发展,尽管融合质量在不断提高,但有利于目标检测任务的算法却相对较少。融合图像虽然充分融合了多模态图像的互补信息,在融合图像的质量上取得了明显进步,但是却并不一定适用于目标检测等下游任务。因此,如何能够将图像融合领域中的突出技术与成果转化到目标检测领域,也是未来的一个重要研究内容。

(3)计算资源的缩减。精度高但实时性差、网络复

表6 不同机制下网络模型对比

Table 6 Comparison of network models under different mechanisms

网络模型	机制	创新	优势	局限性	适用场景
CFR[72]	CNN	循环融合与特征逐步优化	强通用性和兼容性	特征融合不充分,泛化能力差	复杂环境下的行人检测
LRAF-Net[38]	CNN、Transformer注意力机制	通道、空间维度注意力增强,长距离依赖特征融合	特征深度融合与实时性,互补掩膜数据增强	计算复杂度高,依赖数据增强策略	夜间或复杂背景下多目标检测
NRSNet[83]	CNN、光照感知	光照感知动态分配融合权重,3D注意力机制	适应性和泛化能力强,模型轻量化	特征融合和数据集验证不充分	光照条件变化下行人、车辆检测
YOLO-Adaptor[87]	CNN、特征对齐	多模态适配器、特征对比学习损失,未对齐特征处理	轻量化、鲁棒性和泛化能力强,实时性好	非刚性变换适应性有限,小目标检测能力弱	图像未对齐条件下的非小目标检测
CFT[91]	CNN、Transformer	基于Transformer的跨模态深度融合,全局信息整合	精度提升显著,鲁棒性和泛化能力强,效率高	计算开销大,模态不平衡的处理有限	图像对齐条件下的行人、车辆检测
MSANet[93]	CNN、Transformer、TNT	多尺度特征聚合和跨模态信息对齐,局部与全局特征提取	跨模态信息互补与对齐,检测性能好	结构复杂计算开销大,小目标检测能力有限	自动驾驶场景下的行人、车辆检测
GM-DETR[95]	CNN、Transformer	DETR检测框架,两阶段训练策略,跨模态尺度特征融合	检测精度高、实时性好,模型轻量化	光照或天气变化时泛化能力弱,小目标检测弱	图像对齐、弱干扰下多光谱目标检测
MGMF[103]	CNN、Mamba	掩码正则化约束提取跨模态引导信息,状态空间融合	特征深度融合,端到端检测,检测精度高	结构复杂、实时性差,泛化能力验证不足	无人机平台上的可见光-红外车辆检测
RemoteDet-Mamba[106]	CNN、Mamba	四向选择性扫描融合,CNN-Mamba的局部-全局特征提取	小目标可区分性强,检测精度高、实时性好	环境适应性验证不足,泛化能力弱	目标密集条件下的无人机遥感检测
COMO[107]	CNN、Mamba	跨模态Mamba交互,全局与局部扫描和偏移引导融合	泛化能力强,模型参数和计算量低,实时性好	过度依赖高级特征,网络结构复杂	图像未对齐的可见光-红外目标检测

杂且占用计算资源大等问题,严重限制了多模态图像融合检测在无人机、无人驾驶等领域的应用。当前基于Transformer的目标检测网络表现出了令人瞩目的检测精度,但是其巨大的计算消耗却大大限制了其在实际应用中的拓展。此外,融合检测由于其处理双源信息的双流结构,所产生的参数量和计算量也远大于单模态目标检测。因此,如何在保证检测精度的同时有效缩减计算资源,也将是未来研究的主要方向之一。

(4)不同应用场景下的适应性。目标检测的应用需求广泛,例如在低空经济中扮演重要角色的无人机、逐渐走进大众生活的无人智能驾驶和当前军事领域的智能化作战等。因此,如何设计网络结构以适应不同的应用场景也是当前面临的主要难题,例如在无人驾驶时需要重点考虑检测精度和实时性,而在无人机的目标检测时则要考虑网络的轻量化。此外,处于恶劣环境时如雨雪、大雾和海上盐雾与反光等,网络需要充分利用双模态图像的互补优势,自适应的实现对目标的精准定位和分类。因此,如何提高网络在各种应用场景下的鲁棒性,也是未来研究的一个重要方向。

4 结束语

基于深度学习的可见光-红外图像融合的目标检测是目标检测技术发展的一个重要分支,多模态图像可以为目标检测网络提供更加丰富的信息来源,使其在复杂环境下仍能保持较高的检测精度和鲁棒性。本文首先介绍了双模态图像融合检测的研究现状,着重分析了基于不同融合阶段和不同基础模型的检测算法,通过介绍当前主流的研究方向与最新研究成果,为该领域的进一步研究指明方向:一是基于特征级融合的目标检测将继续占据该领域研究的主导地位,同时像素级融合与决策级融合也具有一定的创新空间;二是基于Transformer与基于Mamba的融合检测网络将会是未来研究的主要方向,发展的重点则是以实际开发应用为最终目标,相对于发展数年的Transformer,基于Mamba的研究无疑更具有研究的潜力。

参考文献:

[1] QI Y, HOU W, YANG L Q, et al. GMBox: box-supervised remote sensing images instance segmentation based on multi-scale gradient prior fusion and mask correction[C]//Proceedings of the 11th International Conference on Information Systems and Computing Technology. Piscataway: IEEE, 2023: 268-273.

[2] LI L C, CHEN W, QI J. VB-SOLO: single-stage instance segmentation of overlapping epithelial cells[J]. IEEE Access, 2024, 12: 52555-52564.

[3] WANG X, SUN Z J, CHEHRI A, et al. Deep learning and multi-modal fusion for real-time multi-object tracking: algorithms, challenges, datasets, and comparative study[J]. Information Fusion, 2024, 105: 102247.

[4] DU Z H, JIANG H, YANG X, et al. Deep learning-assisted near-Earth asteroid tracking in astronomical images[J]. Advances in Space Research, 2024, 73(10): 5349-5362.

[5] CHEN H, YAN H Q, YANG X, et al. Efficient adversarial attack strategy against 3D object detection in autonomous driv-

ing systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 16118-16132.

[6] SONG Z Y, LIU L, JIA F Y, et al. Robustness-aware 3D object detection in autonomous driving: a review and outlook[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 15407-15436.

[7] ABOUOUF M, SINGH S, MIZOUNI R, et al. Explainable AI for event and anomaly detection and classification in healthcare monitoring systems[J]. IEEE Internet of Things Journal, 2024, 11(2): 3446-3457.

[8] ZHAO Q, WANG Y, WANG B Y, et al. MSC-AD: a multiscene unsupervised anomaly detection dataset for small defect detection of casting surface[J]. IEEE Transactions on Industrial Informatics, 2024, 20(4): 6041-6052.

[9] WANG Q, GAO S, XIONG L, et al. A casting surface dataset and benchmark for subtle and confusable defect detection in complex contexts[J]. IEEE Sensors Journal, 2024, 24(10): 16721-16733.

[10] 梁礼明, 龙鹏威, 卢宝贺, 等. EHH-YOLOv8s: 一种轻量级的带钢表面缺陷检测算法[J/OL]. 北京航空航天大学学报, 2024: 1-15(2024-08-08)[2025-01-05]. https://kns.cnki.net/KCMS/detail/detail.aspx?filename=BJHK20240806002&dbname=CJFD&dbcode=CJFQ.
LIANG L M, LONG P W, LU B H, et al. EHH-YOLOv8s: a lightweight algorithm for strip surface defect detection[J/OL]. Journal of Beijing University of Aeronautics and Astronautics, 2024: 1-15(2024-08-08)[2025-01-05]. https://kns.cnki.net/KCMS/detail/detail.aspx?filename=BJHK20240806002&dbname=CJFD&dbcode=CJFQ.

[11] 王元喆, 梁腾飞, 曾宇乔, 等. 多光谱目标检测综述[J]. 信息与控制, 2024, 53(3): 287-301.
WANG Y Z, LIANG T F, ZENG Y Q, et al. Overview of multispectral target detection[J]. Information and Control, 2024, 53(3): 287-301.

[12] LUO Y Y, LUO Z Q. Infrared and visible image fusion: methods, datasets, applications, and prospects[J]. Applied Sciences, 2023, 13(19): 10891.

[13] MA W, WANG K, LI J, et al. Infrared and visible image fusion technology and application: a review[J]. Sensors (Basel), 2023, 23(2): 599.

[14] JIAO T Z, GUO C P, FENG X Y, et al. A comprehensive survey on deep learning multi-modal fusion: methods, technologies and applications[J]. Computers, Materials & Continua, 2024, 80(1): 1-35.

[15] WANG Z A, LIAO X H, YUAN J, et al. CDC-YOLOFusion: leveraging cross-scale dynamic convolution fusion for visible-infrared object detection[J]. IEEE Transactions on Intelligent Vehicles, 2024: 1-14.

[16] LEE W Y, JOVANOV L, PHILIPS W. Multimodal pedestrian detection based on cross-modality reference search[J]. IEEE Sensors Journal, 2024, 24(10): 17291-17306.

[17] LI Q, ZHANG C Q, HU Q H, et al. Stabilizing multispectral pedestrian detection with evidential hybrid fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(4): 3017-3029.

[18] JIA X Y, ZHU C, LI M Z, et al. LLVIP: a visible-infrared paired dataset for low-light vision[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE, 2021: 3489-3497.

[19] HUANG N C, LIU J N, MIAO Y Q, et al. Deep learning for visible-infrared cross-modality person re-identification: a comprehensive review[J]. Information Fusion, 2023, 91: 396-411.

[20] LIU J Y, FAN X, JIANG J, et al. Learning a deep multiscale feature ensemble and an edge-attention guidance for image fusion[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(1): 105-119.

[21] ZHANG X, ZHANG X H, WANG J T, et al. TFDet: target-aware fusion for RGB-T pedestrian detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(7): 13276-13290.

[22] 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. Piscataway: IEEE, 2014: 580-587.

[23] WANG S W, LI Y, QIAO S H. ALF-YOLO: enhanced YOLOv8 based on multiscale attention feature fusion for ship detection [J]. Ocean Engineering, 2024, 308: 118233.

[24] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway: IEEE, 2015: 1440-1448.

[25] 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 & Machine Intelligence, 2017, 39(6): 1137-1149.

[26] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 779-788.

[27] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 6517-6525.

[28] REDMON J, FARHADI A. YOLOv3: an incremental improvement[J]. arXiv:1804.02767, 2018.

[29] BOCHKOVSKIY A, WANG C Y, LIAO H. YOLOv4: optimal speed and accuracy of object detection[J]. arXiv:2004.10934, 2020.

[30] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers[C]//Proceedings of the

European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 213-229.
[31] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 6000-6010.
[32] LI C, HEI Y Q, XI L H, et al. GL-DETR: global-to-local transformers for small ship detection in SAR images[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 3461212.
[33] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2024: 16965-16974.
[34] NIE J Y, SUN H, SUN X, et al. Cross-modal feature fusion and interaction strategy for CNN-transformer-based object detection in visual and infrared remote sensing imagery[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.
[35] ALSHEHRI M, OUADOU A, SCOTT G J. Deep transformer-based network deforestation detection in the Brazilian Amazon using sentinel-2 imagery[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.
[36] KEDDOUS F E, LLANZA A, SHVAI N, et al. Vision transformers inference acceleration based on adaptive layer normalization[J]. Neurocomputing, 2024, 610: 128524.
[37] WANG J, LI X, CHEN R F, et al. Infrared and visible image fusion based on co-gradient edge-attention gate network[C]//Proceedings of the 9th International Conference on Control and Robotics Engineering. Piscataway: IEEE, 2024: 339-344.
[38] FU H, WANG S, DUAN P, et al. LRAF-Net: long-range attention fusion network for visible-infrared object detection [J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(10): 13232-13245.
[39] DING R, YANG M, ZHENG N N. Selective transfer learning of cross-modality distillation for monocular 3D object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(10): 9925-9938.
[40] BURT P J. The pyramid as a structure for efficient computation[M]. Cham: Springer, 1984.
[41] YU M, CUI T, LU H Y, et al. VIFNet: an end-to-end visible-infrared fusion network for image dehazing[J]. Neurocomputing, 2024, 599: 128105.
[42] LI X, HE H, SHI J. HDCCT: hybrid densely connected CNN and transformer for infrared and visible image fusion[J]. Electronics, 2024, 13(17): 3470.
[43] CHEN X X, XU S W, HU S H, et al. ACFNet: an adaptive cross-fusion network for infrared and visible image fusion [J]. Pattern Recognition, 2025, 159: 111098.
[44] TANG W, HE F Z, LIU Y. ITFuse: an interactive transformer for infrared and visible image fusion[J]. Pattern Recognition, 2024, 156: 110822.
[45] 张宏钢, 杨海涛, 郑逢杰, 等. 特征级红外与可见光图像融合方法综述[J]. 计算机工程与应用, 2024, 60(18): 17-31.
ZHANG H G, YANG H T, ZHENG F J, et al. Review of feature-level infrared and visible image fusion[J]. Computer Engineering and Applications, 2024, 60(18): 17-31.
[46] LI Z, PAN H, ZHANG K, et al. MambaDFuse: a mamba-based dual-phase model for multi-modality image fusion [J]. arXiv:2404.08406, 2024.
[47] ZHANG X, DEMIRIS Y. Visible and infrared image fusion using deep learning[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2023, 45(8): 10535-10554.
[48] SHOPOVSKA I, JOVANOV L, PHILIPS W. Deep visible and thermal image fusion for enhanced pedestrian visibility [J]. Sensors (Basel), 2019, 19(17): e3727.
[49] LIU J Y, FAN X, HUANG Z B, et al. Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 5792-5801.
[50] HOU Z Q, LI X Y, YANG C, et al. Dual-branch network object detection algorithm based on dual-modality fusion of visible and infrared images[J]. Multimedia Systems, 2024, 30 (6): 333.
[51] HWANG S, PARK J, KIM N, et al. Multispectral pedestrian detection: benchmark dataset and baseline[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2015: 1037-1045.
[52] FREE Teledyne FLIR thermal dataset for algorithm training [EB/OL]. (2018-02.22)[2024-11-21]. https://www.flir.com/oem/adas/adas-dataset-form/.
[53] RAZAKARIVONY S, JURIE F. Vehicle detection in aerial imagery: a small target detection benchmark[J]. Journal of Visual Communication and Image Representation, 2016, 34: 187-203.
[54] SUN Y M, CAO B, ZHU P F, et al. 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.
[55] SONG K C, XUE X T, WEN H W, et al. Misaligned visible-thermal object detection: a drone-based benchmark and baseline[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(11): 7449-7460.
[56] LI C Y, SONG D, TONG R F, et al. Illumination-aware faster R-CNN for robust multispectral pedestrian detection[J]. Pattern Recognition, 2019, 85: 161-171.
[57] YAN C Q, ZHANG H, LI X L, et al. Cross-modality complementary information fusion for multispectral pedestrian detection[J]. Neural Computing and Applications, 2023, 35 (14): 10361-10386.
[58] ZHANG L, LIU Z, ZHU X, et al. Weakly aligned feature

fusion for multimodal object detection[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(3): 4145-4159.

[59] ZENG Y, LIANG T, JIN Y, et al. MMI-Det: exploring multimodal integration for visible and infrared object detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(11): 11198-11213.

[60] SUN Y X, MENG Y Q, WANG Q B, et al. Visible and infrared image fusion for object detection: a survey[C]//Proceedings of the International Conference on Image, Vision and Intelligent Systems , 2024: 236-248.

[61] WAGNER J, FISCHER V, HERMAN M, et al. Multispectral pedestrian detection using deep fusion convolutional neural networks[J]. arXiv:1611.02644, 2016.

[62] PENG R H, LAI J, YANG X T, et al. Camouflaged target detection based on multimodal image input pixel-level fusion [J]. Frontiers of Information Technology & Electronic Engineering, 2024, 25(9): 1226-1239.

[63] SHEN J F, CHEN Y F, LIU Y, et al. ICAFusion: iterative cross-attention guided feature fusion for multispectral object detection [J]. Pattern Recognition, 2024, 145: 109913.

[64] XIE Y M, ZHANG L W, YU X Y, et al. YOLO-MS: multispectral object detection via feature interaction and self-attention guided fusion[J]. IEEE Transactions on Cognitive and Developmental Systems, 2023, 15(4): 2132-2143.

[65] ZHANG Y, YU H, HE Y J, et al. Illumination-guided RGBT object detection with inter- and intra-modality fusion[J]. IEEE Transactions on Instrumentation Measurement, 2023, 72: 3251414.

[66] LI Q, ZHANG C Q, HU Q H, et al. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection[J]. IEEE Transactions on Multimedia, 2023, 25: 3420-3431.

[67] HU Z H, JING Y G, WU G Q. Decision-level fusion detection method of visible and infrared images under low light conditions[J]. EURASIP Journal on Advances in Signal Processing, 2023, 2023(1): 38.

[68] KANG X D, YIN H, DUAN P H. Global local feature fusion network for visible infrared vehicle detection[J]. IEEE Geoscience and Remote Sensing Letters, 2024, 21: 1-5.

[69] YU H Y, YANG H, GAO L R, et al. Hyperspectral image change detection based on gated spectral spatial temporal attention network with spectral similarity filtering[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-13.

[70] SHI M N, LI H T, YAO Q, et al. Vision based nighttime pavement cracks pixel level detection by integrating infrared visible fusion and deep learning[J]. Construction and Building Materials, 2024, 442: 137662.

[71] LIU J J, ZHANG S T, WANG S, et al. Multispectral deep neural networks for pedestrian detection[J]. arXiv:1611.02644, 2016.

[72] ZHANG H, FROMONT E, LEFEVRE S, et al. Multispectral fusion for object detection with cyclic fuse-and-refine blocks[C]//Proceedings of the IEEE International Conference on Image Processing. Piscataway: IEEE, 2020: 276-280.

[73] XIAO X W, WANG B, MIAO L J, et al. Infrared and visible image object detection via focused feature enhancement and cascaded semantic extension[J]. Remote Sensing, 2021, 13(13): 2538.

[74] FENG Y, LUO E B, LU H, et al. Cross-modality feature fusion for night pedestrian detection[J]. Frontiers in Physics, 2024, 12: 1356248.

[75] JADERBERG M, SIMONYAN K, ZISSERMAN A, et al. Spatial transformer networks[J]. arXiv:1506.02025, 2015.

[76] HU J, SHEN L, SUN G. Squeeze-and-excitation networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7132-7141.

[77] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer International Publishing, 2018: 3-19.

[78] YANG Y, XU K X, WANG K Z. Cascaded information enhancement and cross-modal attention feature fusion for multispectral pedestrian detection[J]. Frontiers in Physics, 2023, 11: 1121311.

[79] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2021: 9992-10002.

[80] LI R M, XIANG J J, SUN F X, et al. Multiscale cross-modal homogeneity enhancement and confidence-aware fusion for multispectral pedestrian detection[J]. IEEE Transactions on Multimedia, 2024, 26: 852-863.

[81] HU S J, BONARDI F, BOUCHAFA S, et al. Rethinking self-attention for multispectral object detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2024, 25(11): 16300-16311.

[82] LIU X W, XU X Y, XIE J, et al. FDENet: fusion depth semantics and edge-attention information for multispectral pedestrian detection[J]. IEEE Robotics and Automation Letters, 2024, 9(6): 5441-5448.

[83] 程清华, 鉴海防, 郑帅康, 等. 基于光照感知的红外/可见光融合目标检测[J]. 计算机科学, 2025, 52(2): 173-182.
CHENG Q H, JIAN H F, ZHENG S K, et al. Illumination-aware infrared/visible fusion for object detection[J]. Computer Science, 2025, 52(2): 173-182.

[84] ZHANG L, ZHU X Y, CHEN X Y, et al. Weakly aligned cross-modal learning for multispectral pedestrian detection

[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 5126-5136.
[85] CHEN Y X, GUAN Y, SHAO Z Z. Real-time multispectral pedestrian detection with weakly aligned cross-modal learning [C]//Proceedings of the IEEE International Conference on Real-time Computing and Robotics. Piscataway: IEEE, 2023: 829-834.
[86] TIAN C, ZHOU Z K, HUANG Y Q, et al. Cross-modality proposal-guided feature mining for unregistered RGB-thermal pedestrian detection[J]. IEEE Transactions on Multimedia, 2024, 26: 6449-6461.
[87] FU H L, LIU H H, YUAN J, et al. YOLO-Adaptor: a fast adaptive one-stage detector for non-aligned visible-infrared object detection[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(11): 7070-7083.
[88] ZHOU K L, CHEN L S, CAO X. Improving multispectral pedestrian detection by addressing modality imbalance problems[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer International Publishing, 2020: 787-803.
[89] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv:2010.11929, 2020.
[90] CHEN X, YAN B, ZHU J, et al. High-performance transformer tracking[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2023, 45(7): 8507-8523.
[91] FANG Q, HAN D, WANG Z K. Cross-modality fusion transformer for multispectral object detection[J]. arXiv:2010.11929, 2020.
[92] LEE W Y, JOVANOV L, PHILIPS W. Cross-modality attention and multimodal fusion transformer for pedestrian detection[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer International Publishing, 2023: 608-623.
[93] YOU S, XIE X D, FENG Y J, et al. Multi-scale aggregation transformers for multispectral object detection[J]. IEEE Signal Processing Letters, 2023, 30: 1172-1176.
[94] HAN K, XIAO A, WU E, et al. Transformer in transformer [C]//Advances in Neural Information Processing Systems, 2021: 15908-15919.
[95] XIAO Y M, MENG F M, WU Q B, et al. GM-DETR: generalized muiltispectral detection Transformer with efficient fusion encoder for visible-infrared detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2024: 5541-5549.
[96] GAO H W, WANG Y T, SUN J, et al. Efficient multi-level cross-modal fusion and detection network for infrared and visible image[J]. Alexandria Engineering Journal, 2024, 108: 306-318.
[97] GU A, DAO T. Mamba: linear-time sequence modeling with selective state spaces[J]. arXiv:2312.00752, 2023.
[98] YU W, WANG X. MambaOut: do we really need mamba for vision?[J]. arXiv:2405.07992, 2024.
[99] ZHU L, LIAO B, ZHANG Q, et al. Vision Mamba: efficient visual representation learning with bidirectional state space model[J]. arXiv:2401.09417, 2024.
[100] LIU Y, TIAN Y, ZHAO Y, et al. VMamba: visual state space model[C]//Advances in Neural Information Processing Systems, 2024: 103031-103063.
[101] DONG W, ZHU H, LIN S, et al. Fusion-Mamba for cross-modality object detection[J]. arXiv:2404.09146, 2024.
[102] LIANG J Y, CAO J Z, SUN G L, et al. SwinIR: image restoration using swin transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. Piscataway: IEEE, 2021: 1833-1844.
[103] WANG S M, WANG C P, SHI C Y, et al. Mask-guided Mamba fusion for drone-based visible-infrared vehicle dete-ction[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 3452550.
[104] LI H Y, HU Q, YAO Y, et al. CFMW: cross-modality fusion Mamba for multispectral object detection under adverse weather conditions[J]. arXiv:2404.16302, 2024.
[105] HE X H, CAO K, ZHANG J, et al. Pan-Mamba: effective pan-sharpening with state space model[J]. Information Fusion, 2025, 115: 102779.
[106] REN K J, WU X, XU L M, et al. RemoteDet-Mamba: a hybrid Mamba-CNN network for multi-modal object detection in remote sensing images[J]. arXiv:2410.13532, 2024.
[107] LIU C, MA X, YANG X C, et al. COMO: cross-Mamba interaction and offset-guided fusion for multimodal object detection[J]. arXiv:2412.18076, 2024.
[108] ZHOU M, LI T, QIAO C, et al. DMM: disparity-guided multispectral Mamba for oriented object detection in remote sensing[J]. arXiv:2407.08132, 2024.