计算机科学
COMPUTER SCIENCE

重参数化增强的双模态实时目标检测模型

李允臣, 张睿, 王家宝, 李阳, 王梓祺, 陈瑶

引用本文

李允臣, 张睿, 王家宝, 李阳, 王梓祺, 陈瑶. 重参数化增强的双模态实时目标检测模型[J]. 计算机科学, 2024, 51(9): 162-172.

LI Yunchen, ZHANG Rui, WANG Jiabao, LI Yang, WANG Ziqi, CHEN Yao. Re-parameterization Enhanced Dual-modal Realtime Object Detection Model [J]. Computer Science, 2024, 51(9): 162-172.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230700106

重参数化增强的双模态实时目标检测模型

李允臣　张　睿　王家宝　李　阳　王梓祺　陈　瑶
陆军工程大学指挥控制工程学院　南京 210007
(liyunchen1012@163.com)

摘　要　无人机高空航拍的目标普遍尺寸小、特征弱,而且受复杂天候条件影响大,导致基于可见光或红外单模态图像的目标检测漏检、误检率较高。对此,提出了重参数化增强的双模态实时目标检测模型 DM-YOLO。首先,采用通道拼接的方法融合可见光和红外图像,以极低的成本融合双模态图像的互补信息。其次,提出更加高效的重参数化模块并基于此构建了更加强大的骨干网 RepCSPDarkNet,有效增强了骨干网对双模态图像的特征提取能力。然后,提出了多层次特征融合模块,通过多感受野卷积和注意力机制融合弱小目标的多尺度特征信息,增强了弱小目标的多尺度特征表示。最后,删除了对弱小目标检测基本不起作用的特征金字塔深层检测层,在检测精度保持不变的情况下,减小了模型规模。实验结果表明,在大规模的双模态图像数据集 DroneVehicle 上,DM-YOLO 的检测精度比基准 YOLOv5s 高出 2.45%,且优于规模相当的 YOLOv6 和 YOLOv7 模型,有效提高了复杂光照条件下目标检测的准确性和鲁棒性,同时检测速度达到 82 FPS,可满足实时检测的需求。

关键词:重参数化;双模态;实时目标检测;多尺度特征;注意力机制
中图分类号　TP391

Re-parameterization Enhanced Dual-modal Realtime Object Detection Model

LI Yunchen,ZHANG Rui,WANG Jiabao,LI Yang,WANG Ziqi and CHEN Yao
College of Command and Control Engineering,Army Engineering University of PLA,Nanjing 210007,China

Abstract　The objects captured by drones at high altitudes are generally small and have weak features,and they are greatly affected by complex weather conditions. Object detection based on visible or infrared images often has high rates of missed detection and false detection. To address this problem,this paper proposes a dual-modal realtime object detection model DM-YOLO with reparameterization enhancement. Firstly,the visible and infrared images are effectively fused by channel concatenation,which makes efficient use of the complementary information in the dual-modal images at a very low cost. Secondly,a more efficient reparameterization module is proposed and a more powerful backbone network RepCSPDarkNet is constructed based on it,which effectively improves the feature extraction capability of the backbone network for dual-modal images. Then,a multi-level feature fusion module is proposed to enhance the multiscale feature representation of weak and small objects by fusing multi-scale feature information of weak and small objects with multi-receptive field dilated convolution and attention mechanism. Finally,the deep feature layer of the feature pyramid is removed,which reduces the model size while maintaining the detection accuracy. Experimental results on the large-scale dual-modal image dataset DroneVehicle show that,the detection accuracy of DM-YOLO is 2.45% higher than that of the baseline YOLOv5s,and is better than that of the YOLOv6 and YOLOv7 models. Furthermore,it effectively improves the accuracy and robustness of object detection under complex weather conditions,while achieving a detection speed of 82 frames per second,which can meet the requirements of realtime detection.

Keywords　Reparameterization,Dual modality,Real-time object detection,Multiscale features,Attention mechanism

1　引言

目标检测是计算机视觉领域的一个重要分支,在自动驾驶、交通监控、应急救援等领域都有非常重要的应用[1-3]。然而,在无人机高空航拍条件下,目标普遍尺寸小、特征弱,而且受复杂天候条件的影响大,目标检测面临较大挑战。

近年来,基于深度学习的目标检测算法由于精度高、速度快,逐渐成为目标检测领域的主流算法。基于深度学习的目标检测算法主要分为两类:一类是以 R-CNN(Region-based Convolutional Neural Network)系列[4-8]为代表的两阶段目标

到稿日期:2023-07-17　返修日期:2023-11-06
基金项目:江苏省高校自然科学研究基金(BK20200581)
This work was supported by the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province,China(BK20200581).
通信作者:张睿(Lydiazhang09@163.com)

检测算法；另一类是以 YOLO（You Only Look Once）系列[9-15]和 SSD（Single Shot Multi-box Detector）系列[16-17]为代表的单阶段目标检测算法。单阶段目标检测算法在检测速度方面具有明显的优势，可进行实时检测，在工业和生活领域应用广泛。

目前，大多数目标检测算法均基于可见光或红外单模态图像进行检测，无法有效适应复杂条件下的目标检测任务。其中，基于可见光图像的目标检测受光照条件影响较大，在夜晚、雨雾等低光照度条件下，图像质量较差，目标模糊不清，甚至完全不可见，如图 1 第一行（c）—（f）所示。而基于红外图像的目标检测虽然不受光照条件影响，但红外图像分辨率较低（通常为 640×512 像素以下），而且缺乏颜色信息，局部纹理细节信息不足，背景噪声较多，目标整体特征较弱，如图 1 第二行所示。

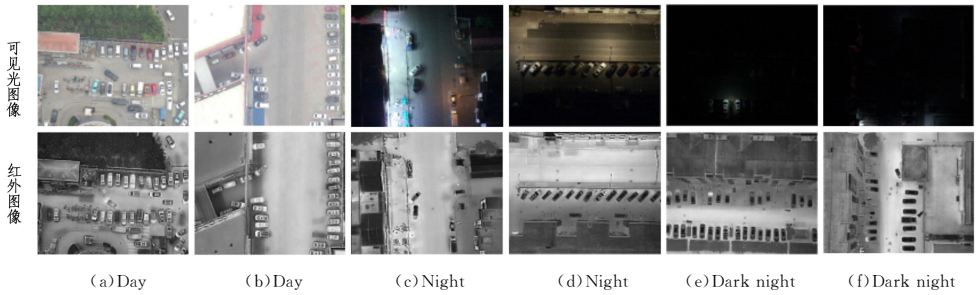


图 1　不同光照条件下无人机航拍图像

Fig. 1　Aerial images under different lighting conditions

与日常生活场景下的目标检测相比，无人机高空航拍条件下的目标检测不仅面临天候因素的干扰，还面临着更多其他的挑战：1）空中航拍的目标普遍尺寸小，局部细节特征不明显，整体特征更弱；2）在空中航拍视角下，拍摄采集的目标为俯视图，相近类别的物体外形相似，特征差异小；3）航拍图像的背景区域大，背景干扰因素更多，进一步增大了检测的难度。

针对以上问题，通过融合可见光和红外双模态图像综合利用两者的互补特征信息，不仅有利于克服夜暗、雨雾等复杂天候条件的不利影响，增强目标检测的鲁棒性，而且可以增强弱小目标的特征表示，提高弱小目标检测的精度。此外，增强神经网络对弱小目标的特征提取能力，抑制背景噪声的干扰，也是提高弱小目标检测精度的有效手段。

本文致力于构建一个可适应复杂天候条件的无人机载实时目标检测模型。因此，本文以轻量级的 YOLOv5s 为基准进行改进，提出了重参数化增强的双模态实时目标检测模型 DM-YOLO（Dual-Modal YOLO）。本文的主要工作如下：

1）提出了一个轻量级的融合可见光/红外双模态图像的实时目标检测模型，不仅能够适应复杂天候条件下的弱小目标检测任务，而且参数量和计算量都较小。

2）提出了更加高效的重参数化模块 ERepBottleneck（Efficient Reparameterized Bottleneck），并基于此构建了更强大的骨干网 RepCSPDarkNet，有效增强了对弱小目标的特征提取能力。

3）针对特征金字塔分辨率不同的特点，提出了多层次特征融合模块 MLFFM（Multi-level Feature Fusion Module），通过多感受野卷积和注意力机制有效融合了弱小目标的多尺度特征信息，增强了弱小目标的多尺度特征表示。

4）在大规模的双模态图像数据集 DroneVehicle 上，所提 DM-YOLO 模型检测精度 mAP_{50} 达到 77.93%，比基准 YOLOv5s 高出 2.45%，并优于规模相当的 YOLOv6 和 YOLOv7 模型。

2　相关工作

2.1　双模态图像融合检测算法

双模态图像融合检测算法通过融合不同模态图像的互补信息，可增强目标的特征表示，从而提高目标检测的鲁棒性和准确性。目前，双模态图像融合检测算法可分为像素级、特征级和决策级。像素级融合检测算法通常先设法将成对的可见光和红外图像融合为一张图像，再对融合后的图像进行目标检测。Wu 等[18]基于深度神经网络设计了梯度残差模块对可见光和红外图像进行融合，相比传统融合算法，保留了更多纹理细节信息并增强了目标亮度，提高了检测精度。Liu 等[19]先利用 GAN 网络[20]的生成器融合可见光和红外图像，而后用两个鉴别器分别鉴别可见光图像的纹理信息和红外图像的目标信息，从而生成高质量的融合图像用于目标检测任务。像素级融合检测方法虽然没有增加检测成本，但必须先训练图像融合网络，而后才能对融合后的图像进行检测，因此，其无法进行端到端的实时检测。特征级融合检测方法通常先分别提取双模态图像的各自特征，再利用融合后的特征进行检测。Geng 等[21]将成对的可见光和红外双模态图像分别送入 Faster R-CNN 骨干网提取特征，将 stage 4 的特征图拼接后，一同输入 stage 5 进行特征融合。Zhou 等[22]将可见光和红外双模态图像分别送入骨干网提取特征后，先构建单模态图像的特征金字塔，而后利用可见光图像的照明条件信息和红外图像的温度信息获取双模态特征的注意力权重，对双模态特征进行注意力加权融合。决策级融合检测方法通常先对可见光图像和红外图像分别进行目标检测，而后再对检测结果进行融合。Chen 等[23]利用贝叶斯规则和跨模态的条件独立性假设，推导提出了一种基于概率集成的决策级融合检测方法。Sun 等[24]同时采用特征级融合和决策级融合策略，提出了 UA-CMDet（Uncertainty-Aware Cross-Modality Vehicle Detection），先利用可见光分支、红外分支以及双模态特征融合分支分别检测，而后提出不确定感知模块将 3 个分支的检测结果加权融合。特征级和决策级融合检测方法均采用并行的网络分支分别提取

各模态特征，虽然能更好地提取各模态图像特征，但参数量、计算量大，检测速度慢。本文采用双模态图像通道拼接融合的方法，既能较好地利用双模态图像的互补信息，又避免了多分支网络参数量、计算量大的问题，可更好地满足实时检测的需求。

2.2 重参数化方法

在深度神经网络发展早期，所有的网络结构设计都采用直连式，但随着网络深度的增加，直连式的网络结构会带来梯度消失的问题，导致模型难以收敛。He 等[25]于 2012 年提出了带有残差连接的 Resnet，有效解决了深度神经网络梯度消失的难题。虽然多分支的神经网络模型性能通常优于直连式的网络模型，但多分支网络训练时显存消耗大，推理速度较慢。对此，Ding 等[26]提出了结构重参数化方法和 RepVGGBlock 模块。基于 RepVGGBlock 构建的 RepVGG 模型将训练与推理解耦，训练时采用多分支结构以获取更好的性能，推理时将多分支结构等效转换为单分支结构，从而在保持模型性能不变的情况下，加快了推理速度。Ding 等[27]在重参数化方法的基础上，借鉴 Inception 的多分支结构设计思想，提出了 DBB(Diverse Branch Block)模块，设计了可等效转化为一个 $K\times K$ 卷积模块的 6 种不同的重参数化模块，进一步扩展了重参数化方法。2022 年苹果公司利用重参数化模块 RepVGGBlock 设计了适用于手机端的轻量级模型 MobileOne[28]，其检测精度和速度超过了 MobileNet[29-30]，ShuffleNet[31-33]等众多轻量级模型。2022 年，美团公司提出的 YOLOv6[14]以及 Wang 等[15]提出的 YOLOv7 等模型也采用了重参数化模块 RepVGGBlock，在不增加模型参数量、计算量的情况下提高了目标检测精度。Chen 等[34]将重参数化方法应用于 GhostNet，提出了高效的 RepGhost 模块，在移动设备上取得了较好效果。相较于以往的重参数化模块，本文改进后的重参数化模块能更好地保留多分支网络特征图的多样性，增强目标的特征表示。

2.3 注意力机制

人类视觉系统在观察物体时，会将注意力聚焦于某些关键区域，而忽略大部分的背景区域，从而更加快速地获取有效信息。深度学习中的注意力机制通过模仿人类视觉的注意力机制，对重要的特征图通道或位置区域赋予更大的权重，从而获得特征图的通道或空间位置注意力。Hu 等[35]提出了 SE(Squeeze and Excitation)注意力模块，通过全局平均池化操作将特征图信息压缩，并通过全连接层学习得到各通道权重。但 SE 注意力模块只考虑了特征图通道的重要程度，没有考虑空间位置的重要程度。Woo 等[36]提出 CBAM 模块(Convolutional Block Attention Module)，在通道注意力的基础上，增加了空间注意力。Hou 等[37]提出坐标注意力 CA(Coordinate Attention)模块，从水平和垂直两个方向建立特征信息的远程注意力依赖关系。Zhang 等[38]将多尺寸卷积与注意力机制相结合提出了金字塔挤压注意力 PSA(Pyramid Squeeze Attention)模块，其采用 3×3，5×5，7×7，9×9 这 4 种不同尺寸的卷积核来获取不同尺度的局部特征，并引入 SE 通道注意力对不同尺度卷积核获取的特征图进行通道加权，在分类、目标检测等任务中都获得了较好的效果。而本文提出的多尺度特征注意力 MFAM(Multiscale Feature Attention Module)模块进一步引入了更细粒度的空间像素注意力，能够更好地聚焦弱小目标特征，对弱小目标的检测更加有效。

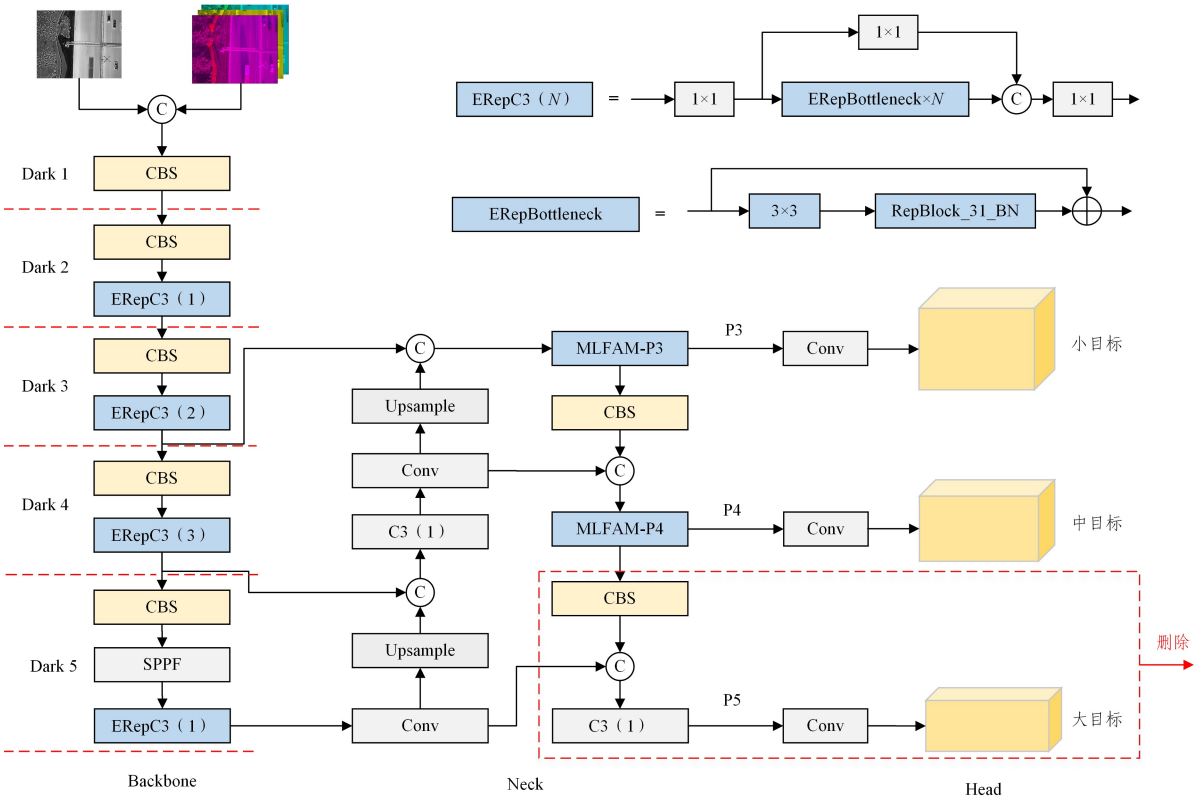


图 2 DM-YOLO 模型整体结构

Fig. 2 Overall structure of DM-YOLO model

3 本文方法

本文提出的 DM-YOLO 模型整体结构如图 2 所示。首先，将可见光图像与红外图像拼接融合后输入模型。然后将重参数化方法改进的 ERepC3 模块替换 YOLOv5s 骨干网 CSPDarkNet 中的 C3 模块，以增强骨干网的特征提取能力。ERepC3 模块主要由 1×1 卷积和 ERepBottleneck 重参数化模块组成(具体见第 3.2.3 节)。接着，将 Neck 部分 P3 和 P4 层的 C3 模块替换为本文提出的多层次特征融合模块 MLFAM-P3 和 MLFAM-P4(具体见第 3.3.2 节)，以更有效地融合不同层次特征图的多尺度特征信息。最后，删除了特征金字塔中用于检测大目标的 P5 检测层，在不降低模型性能的情况下，减少了模型参数量。

3.1 双模态图像拼接融合

为了尽量减少图像融合的参数量和计算量，并实现端到端的训练和检测，本文采取了双模态图像通道拼接融合的方法。首先，将 RGB 图像和 IR 图像归一化至[0，1]区间。然后，将 RGB 图像的 3 通道信息与 IR 图像的单通道信息拼接融合，具体可表示为：

$$X=Concat([R,G,B],[I])=[R,G,B,I] \tag{1}$$

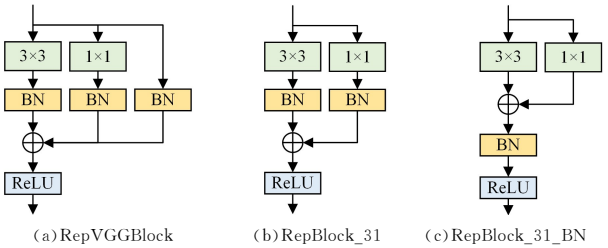
其中，$concat(\cdot)$表示通道拼接操作，$[R,G,B]$表示可见光图像的 3 通道信息，$[I]$表示红外图像的单通道信息，X 表示拼接融合后的图像信息。将双模态图像信息 X 输入骨干网，神经网络通过卷积操作可同时提取可见光与红外图像的特征信息，并对双模态特征信息进行交互融合。

3.2 重参数化方法增强的骨干网

目标检测的性能与骨干网的性能密切相关，YOLOv5s 的骨干网 CSPDarkNet 虽然是一个设计良好的网络架构，但仍存在改进的空间。虽然在骨干网中添加更多的卷积分支可以获得更好的性能，但多分支结构也会增加显存负担和推理时间。为了在不增加推理成本的条件下提升模型性能，本文采用重参数化技术对 CSPDarkNet 进行改进，提出了 RepCSPDarkNet。

3.2.1 基本重参数化模块及其变体

基本的重参数化模块 RepVGGBlock 如图 3(a)所示，其在训练时由 3×3 卷积分支、1×1 卷积分支和恒等映射分支组成，每个分支中均有 BN(Batch Normalization)层，3 分支的输出相加后，再采用 ReLU 函数激活；推理时，先将 BN 层等效并入卷积层，再将 3×3 卷积分支、1×1 卷积分支和恒等映射分支等效合并为 1 层 3×3 卷积分支，在推理时仅有 1 层 3×3 卷积，从而减少了推理成本。



(a)RepVGGBlock (b)RepBlock_31 (c)RepBlock_31_BN

图 3 基本重参数化模块及其变体

Fig. 3 Basic reparameterized module and its variants

RepVGGBlock 模块最初被应用于直连式的 VGG 网络，如果将 RepVGGBlock 模块应用于带有残差连接的网络，则恒等映射分支可能会破坏特征图的多样性，不利于模型性能的提升。因此，Wang 等[15]建议在带有残差连接的网络中去除恒等映射分支，得到重参数化模块的变体 RepBlock_31，如图 3(b)所示。

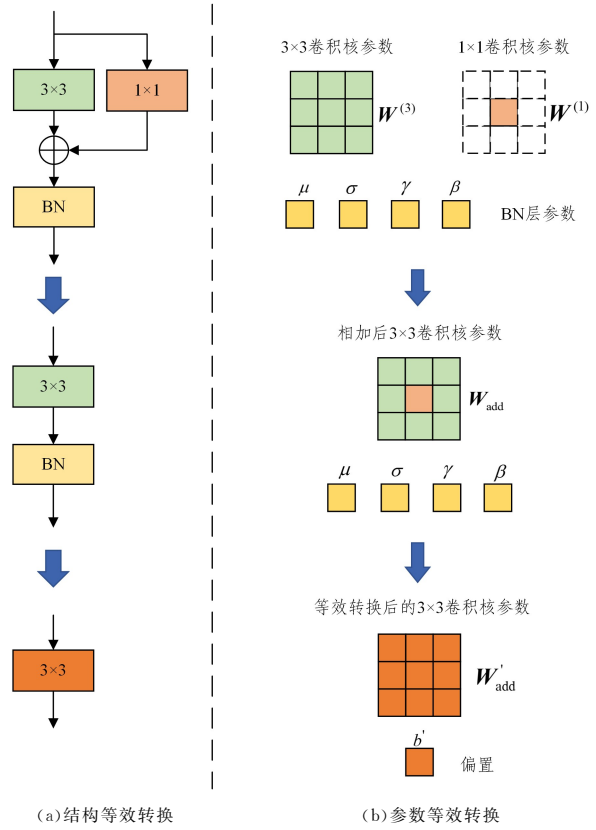
批量归一化层 BN 的作用是将输出的数据分布归一化为均值为 0 且方差为 1 的正态分布，其公式可表达为：

$$Y=BN(X)=(X-\mu)\frac{\gamma}{\sqrt{\sigma^2+\varepsilon}}+\beta \tag{2}$$

其中，X 为输入的数据；μ 和 σ 为批次数据的平均值和方差；γ 为尺度缩放因子，β 为平移因子，γ 和 β 可在训练过程中学习得到；ε 是为避免除数为 0 而添加的一个极小正数。BN 层虽有利于模型收敛，并加快模型训练速度，但不同分支的特征图经过 BN 层处理后，差异性会被削弱。对此，本文先将各分支输出的特征图相加，再经过 BN 层进行归一化处理，从而得到重参数化模块变体 RepBlock_31_BN，如图 3(c)所示。

3.2.2 重参数化模块的等效转换

本文改进的重参数化模块 RepBlock_31_BN 的重参数化等效转换过程如图 4 所示。由于等效转换的过程不涉及 ReLU 激活函数，因此图 4 中未显示 ReLU 函数。



(a)结构等效转换 (b)参数等效转换

图 4 RepBlock_31_BN 模块的重参数化转换过程

Fig. 4 Re-parameterization process of RepBlock_31_BN module

第一步 将 3×3 卷积层和 1×1 卷积层的卷积核权重进行合并。用 $\boldsymbol{F}_{in}\in\boldsymbol{R}^{N\times C_{in}\times H\times W}$ 表示输入的特征图，$\boldsymbol{W}^{(3)}\in\boldsymbol{R}^{C_{out}\times C_{in}\times 3\times 3}$ 表示 3×3 卷积层的权重，$\boldsymbol{F}^{(3)}\in\boldsymbol{R}^{N\times C_{out}\times H\times W}$ 表示 3×3 卷积层输出的特征图；$\boldsymbol{W}^{(1)}\in\boldsymbol{R}^{C_{out}\times C_{in}\times 1\times 1}$ 表示 1×1 卷积

层的权重，$\boldsymbol{F}^{(1)} \in \boldsymbol{R}^{N\times C_{out}\times H\times W}$ 表示 1×1 卷积层输出的特征图；$\boldsymbol{F}_{out} \in \boldsymbol{R}^{N\times C_{out}\times H\times W}$ 表示经过 BN 层处理后最终输出的特征图。其中，N 表示批次输入的图像数量，C_{in} 表示输入的特征图通道数，C_{out} 表示输出的特征图通道数，H 和 W 表示特征图的高和宽。重参数化模块在训练时可表示为：

$$\boldsymbol{F}_{out}=BN(\boldsymbol{F}_{in}*\boldsymbol{W}^{(3)}+\boldsymbol{F}_{in}*\boldsymbol{W}^{(1)})=BN(\boldsymbol{F}^{(3)}+\boldsymbol{F}^{(1)}) \tag{3}$$

重参数化等效转换过程可表示为：

$$\boldsymbol{F}_{out}=BN(\boldsymbol{F}_{in}*(\boldsymbol{W}^{(3)}+\boldsymbol{W}^{(1)}))=BN(\boldsymbol{F}_{in}*\boldsymbol{W}_{add}) \tag{4}$$

其中，$\boldsymbol{W}_{add} \in \boldsymbol{R}^{Cout\times C_{in}\times 3\times 3}$ 表示 $\boldsymbol{W}^{(3)}$ 与 $\boldsymbol{W}^{(1)}$ 相加之和。由于 $\boldsymbol{W}^{(3)}$ 和 $\boldsymbol{W}^{(1)}$ 的尺寸并不匹配，因此，在具体实现时需要先对 $\boldsymbol{W}^{(1)}$ 进行 padding 操作，将其四周填充 0 后，再加入 $\boldsymbol{W}^{(3)}$ 的中心位置，如图 4 所示。$\boldsymbol{W}_{add}$ 可等效替换 3×3 卷积层和 1×1 卷积层，式(1)与式(2)的输出结果完全相同，从而将双分支结构等效转换为单分支结构。

第二步，将 BN 层与卷积层合并。

$$\begin{aligned} BN(\boldsymbol{F}_{in}*\boldsymbol{W}_{add}) &= (\boldsymbol{F}_{in}*\boldsymbol{W}_{add}-u)\frac{\gamma}{\sigma}+\beta \\ &= \boldsymbol{F}_{in}*\boldsymbol{W}_{add}\frac{\gamma}{\sigma}-\mu\frac{\gamma}{\sigma}+\beta \\ &= \boldsymbol{F}_{in}*\boldsymbol{W}'_{add}+b' \end{aligned} \tag{5}$$

其中，$\boldsymbol{W}'_{add}=\boldsymbol{W}_{add}\frac{\gamma}{\sigma}$ 与 $b'=-\frac{u\gamma}{\sigma}+\beta$，分别表示等效转换后得到的 3×3 卷积层的权重和偏置。

经过以上变换，即可将 RepBlock_31_BN 模块等效转换为 1 层 3×3 卷积层和 1 层 ReLU 激活函数。

3.2.3 重参数化增强的骨干网

CSPDarkNet 骨干网主要由 C3 模块组成，而 C3 模块由 3 个 1×1 卷积和若干个 Bottleneck 模块组成。Bottleneck 模块作为骨干网的主体结构，对于提取图像特征至关重要。Bottleneck 模块为 1×1 卷积、3×3 卷积和残差连接组成的残差模块，如图 5(a)所示。

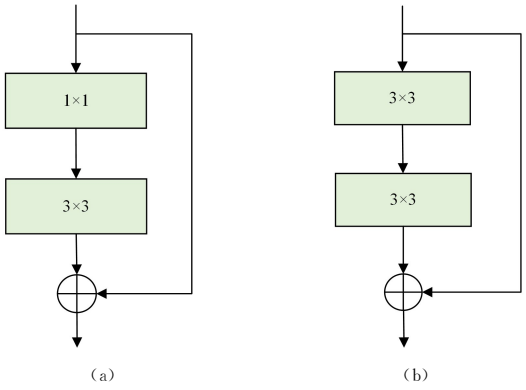


图 5 瓶颈模块及其改进

Fig. 5 Bottleneck module and its improvement

由于 3×3 卷积具有比 1×1 卷积更强的特征提取能力，因此，本文将 1×1 卷积替换为 3×3 卷积，如图 5(b)所示。虽然将 1×1 卷积替换为 3×3 卷积会增大模型的参数量和计算量，但现代 GPU 和 CPU 上的计算库，如 NVDIA 的 cuDNN 和 Intel 的 MKL 等，均对 3×3 卷积运算进行了专门的优化，3×3 卷积具有更快的处理速度。因此，采用 3×3 卷积替换 1×1 卷积并不会显著增加模型推理时间。

为了进一步增强瓶颈结构的特征提取能力，本文将普通 3×3 卷积替换为重参数化模块。利用 3.2.1 节所述的 3 种基本的重参数化模块，本文设计了 9 种不同结构的重参数化瓶颈模块，如图 6 所示。图 6(a)表示用 RepVGGBlock 模块替换瓶颈结构中的两个 3×3 卷积，图 6(b)和图 6(c)表示用 RepVGGBlock 模块替换瓶颈结构中的 1 个 3×3 卷积的两种情形。与 RepVGGBlock 模块类似，如果用 RepBlock_31 模块替换瓶颈结构中的 3×3 卷积，则有图 6(d)－图 6(f) 3 种情形；如果用 RepBlock_31_BN 模块替换瓶颈结构中的 3×3 卷积，则有图 6(g)－图 6(i) 3 种情形。

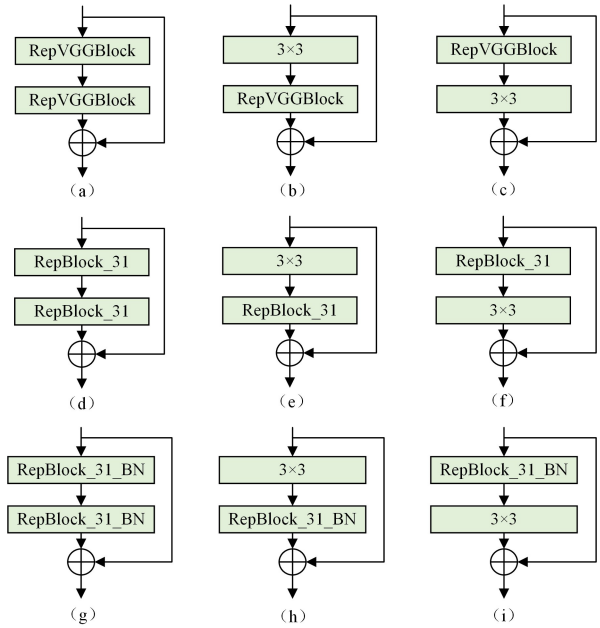


图 6 重参数化瓶颈模块

Fig. 6 Reparameterized bottleneck modules

针对这 9 种瓶颈结构，本文在第 5.1 节进行了实验对比，并最终采用(h)模块，称为 ERepBottleneck。将 ERepBottleneck 模块替换原 C3 模块中的 Bottleneck 后，即得到 ERepC3(Efficient Reparameterized C3)。将骨干网 CSPDarkNet 中 Dark2-Dark5 的 C3 模块替换为 ERepC3 模块，即得到本文所提骨干网 RepCSPDarkNet。RepCSPDarkNet 在训练时有更多的梯度传播路径，可提取更加丰富的目标特征，性能更好；在推理时，其多分支路径可等效合并为单分支路径，从而在模型性能不变的情况下加快推理速度。

3.3 多尺度特征融合

对于目标检测而言，一方面不同大小的目标的有效感受野并不相同，另一方面同一目标需要不同尺度的局部特征共同表示。因此，通过提取目标丰富的多尺度特征，可以有效提高目标检测的精度。由于 YOLOv5 Neck 部分的特征金字塔既有深层特征图的语义信息，又有浅层特征图的细节信息，包含了丰富的多尺度特征，因此，本文提出了多层次特征融合模块用于融合特征金字塔的多尺度特征信息，以增强目标的特征表示。

3.3.1 多尺度特征注意力模块

原 PSA 注意力模块采用 3×3，5×5，7×7，9×9 的

多尺寸卷积虽然有利于提取目标的多尺度特征，但是也存在一些不足：一是大尺寸的卷积核在应用于弱小目标检测时，可能引入更多的噪声，造成目标检测性能下降；二是其仅采用了通道注意力，而忽视了更细粒度的空间注意力，不利于聚焦弱小目标的空间位置信息。对此，本文提出了多尺度特征注意力模块 MFAM，如图 7 所示。MFAM 模块一方面将 PSA 注意力模块的卷积核尺寸适当缩小，以适应无人机航拍条件下对弱小目标的检测；另一方面，在其通道注意力的基础上，进一步添加了更细粒度的空间像素注意力，从通道和像素两个维度聚焦目标的不同尺度特征和关键像素点特征，从而更有利于提取弱小目标的特征信息。

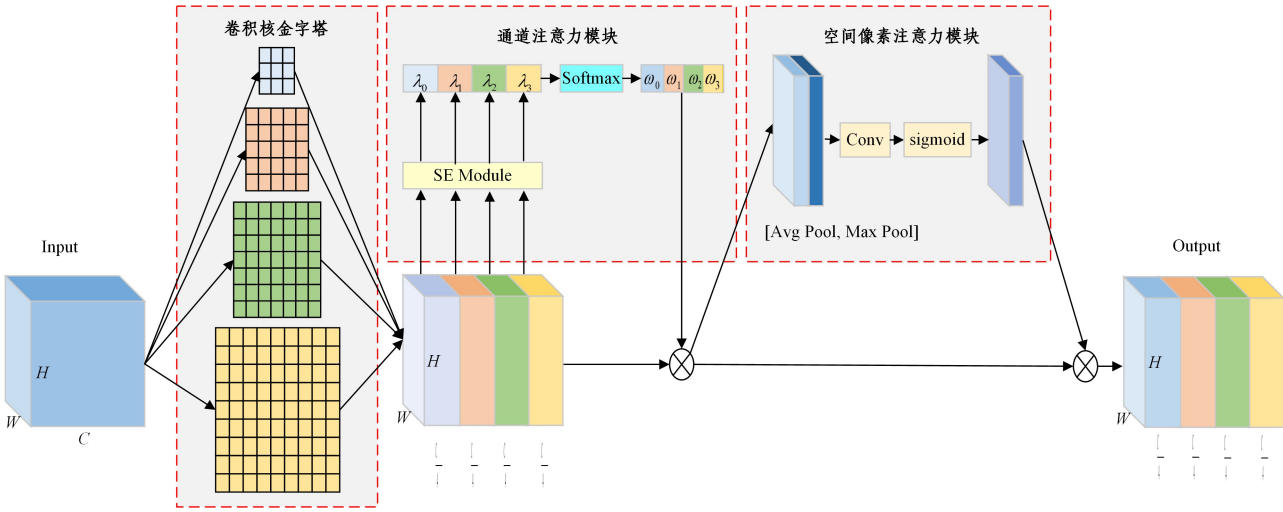


图 7 多尺度特征注意力模块

Fig. 7 Multiscale feature attention module

首先，对于输入的特征图 $\boldsymbol{F}_{\text{input}} \in \boldsymbol{R}^{N\times C\times H\times W}$，MFAM 模块采用卷积核尺寸 $K=[K_1,K_2,K_3,K_4]$，分组 $G=[G_1,G_2,G_3,G_4]$ 的多尺度卷积分别提取目标不同尺度的特征，得到特征图 $\boldsymbol{F}_i \in \boldsymbol{R}^{N\times \frac{C}{4}\times H\times W}$。其中，$N$ 为批次图像的数量，C 为特征图通道数，H 和 W 为特征图的高和宽。接着，采用 SE 注意力模块分别获取 $\boldsymbol{F}_i$ 不同通道的注意力权重 $\boldsymbol{\lambda}_i \in \boldsymbol{R}^{N\times \frac{C}{4}\times 1\times 1}$，并用 Softmax 函数对各通道注意力权重进行归一化校准，获得不同尺度特征之间的全局依赖关系 $\boldsymbol{\omega}_i \in \boldsymbol{R}^{N\times \frac{C}{4}\times 1\times 1}$。将 $\boldsymbol{F}_i$ 与 $\boldsymbol{\omega}_i$ 逐通道相乘后拼接，得到特征图 $\boldsymbol{F}_C \in \boldsymbol{R}^{N\times C\times H\times W}$。该过程可表示为：

$$\boldsymbol{F}_i = Conv_i(\boldsymbol{F}_{\text{input}}), i=0,1,2,3 \tag{6}$$

$$\boldsymbol{\lambda}_i = SEWeight(\boldsymbol{F}_i), i=0,1,2,3 \tag{7}$$

$$\boldsymbol{\omega}_i = \text{Softmax}(\boldsymbol{\lambda}_i) = \frac{\exp(\boldsymbol{\lambda}_i)}{\sum_{i=0}^{3}\exp(\boldsymbol{\lambda}_i)} \tag{8}$$

$$\boldsymbol{F}_C = Concat(\boldsymbol{F}_i \odot \boldsymbol{\omega}_i), i=0,1,2,3 \tag{9}$$

其中，$Conv_i(\cdot)$ 表示不同尺度卷积核的卷积操作，$SEWeight(\cdot)$ 表示 SE 注意力模块，Softmax(·)表示 softmax 函数，$Concat$(·)表示特征图拼接操作，⊙表示逐通道相乘。

对于通道注意力加权后的特征图 $\boldsymbol{F}_C$，通过全局平均池化操作和全局最大池化操作，得到 $\boldsymbol{F}_{\text{avg}} \in \boldsymbol{R}^{N\times 1\times H\times W}$ 和 $\boldsymbol{F}_{\max} \in \boldsymbol{R}^{N\times 1\times H\times W}$。将 $\boldsymbol{F}_{\text{avg}}$ 和 $\boldsymbol{F}_{\max}$ 拼接后，采用 3×3 的普通卷积将通道数压缩为 1，并经过 Sigmoid 函数激活，即得到空间像素注意力权重 $\boldsymbol{S} \in \boldsymbol{R}^{N\times 1\times H\times W}$。将 $\boldsymbol{S}$ 与特征图 $\boldsymbol{F}_C$ 的各通道逐像素相乘，即得到最终输出特征图 $\boldsymbol{F}_{\text{out}} \in \boldsymbol{R}^{N\times C\times H\times W}$。该过程可表示为：

$$\boldsymbol{S} = \sigma(Conv([\boldsymbol{F}_{\text{avg}};\boldsymbol{F}_{\max}])) \tag{10}$$

$$\boldsymbol{F}_{\text{out}} = \boldsymbol{F}_C \otimes \boldsymbol{S} \tag{11}$$

其中，σ 为 Sigmoid 激活函数，$Conv$ 为 3×3 卷积，⊗表示逐像素相乘。

3.3.2 多层次特征融合模块

YOLOv5s 在特征金字塔的 Neck 部分，将深层特征图与浅层特征图拼接后，利用 C3 模块进行融合，不同层次的特征图融合后组成特征金字塔。特征金字塔中既包含了深层特征图的语义信息，还包含了浅层特征图的细节信息。但是 C3 模块仅有 1×1 和 3×3 两种尺寸的卷积核，感受野较为有限，对目标的多尺度特征提取不够充分。为了更好地获取不同层次特征图的多尺度特征信息，本文设计了多层次特征融合模块 MLFFM，如图 8 所示。MLFFM 模块采用 3×3 卷积替换瓶颈结构中的 1×1 卷积，以增强模块的非线性表示能力，并采用多尺度的 MFAM 模块替换瓶颈结构中的 3×3 卷积，从而更有利于提取多层次特征图中的多尺度特征信息。

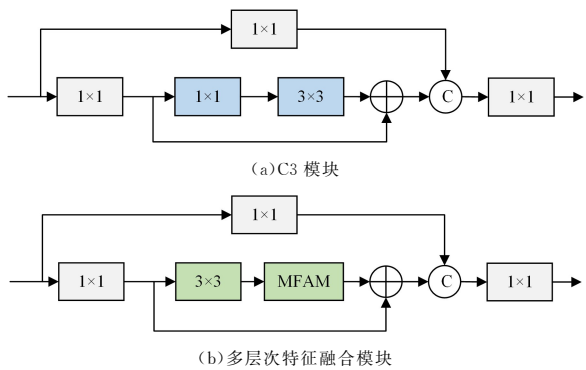


图 8 C3 模块与多层次特征融合模块

Fig. 8 C3 module and multi-level feature fusion module

特征金字塔中的 P3,P4,P5 特征层的分辨率是逐步降低的,其特征图尺寸分别是原始输入图像的 1/8,1/16 和 1/32。由于无人机航拍图像中的目标普遍尺寸小、特征弱,具有较高分辨率的 P3 和 P4 特征层保留了更多的弱小目标细节信息,对于弱小目标检测更加有效。因此,本文设计了与之相匹配的不同尺度的 MLFFM 模块,称为 MLFFM-P3,MLFFM-P4。其中,MLFFM-P3 中的 MFAM 模块采用的卷积核尺寸为 $K=[3,3,5,7]$;MLFFM-P4 模块则进一步缩小卷积核尺寸,取 $K=[1,3,3,5]$。为了缓解 5×5 和 7×7 大尺寸卷积核带来的参数量、计算量大的问题,采用分组卷积策略,将分组数分别设为 2 和 4。

由于 DroneVehicle 数据集中绝大部分目标为弱小目标,而特征金字塔的深层特征图 P5 特征层相较于原始输入图像经过了 5 次下采样,特征图尺寸变为原来的 1/32,过度的下采样导致弱小目标的特征信息损失较多,不利于弱小目标的检测。因此,本文在 Neck 部分删除了 P5 检测层,仅保留 P3 和 P4 检测层。对于删除 P5 检测层的影响,本文将在第 5.4 节进行实验验证。

4 实验设置

4.1 实验环境及训练设置

本文采用 Ubuntu18.04 操作系统,CPU 为 2 块 Intel XEon 5218R,内存 64 GB,显卡为 2 块 NVIDIA GeForce RTX3090,显存 48 GB,CUDA 版本 11.3,cuDNN 版本 8.2.1,PyTorch 版本 1.12.0。

实验超参数采用 YOLOv5s 默认设置,采用 SGD 优化器,初始学习率 0.01,学习动量为 0.937,权重衰减系数为 0.0005,batch size 设为 8,迭代 100 个 epoch。训练时,对图像进行 0.5～1.5 倍的随机缩放,同时采用水平翻转、平移、马赛克增强等数据增强操作。

4.2 评价指标

目标检测常用的评价指标有精确率 P(Precision)、召回率 R(Recall)、平均精度均值 mAP(Mean Average Precision)和检测速度 FPS(Frames Per Second)。计算公式为:

$$P=\frac{TP}{TP+FP} \tag{12}$$

$$R=\frac{TP}{TP+FN} \tag{13}$$

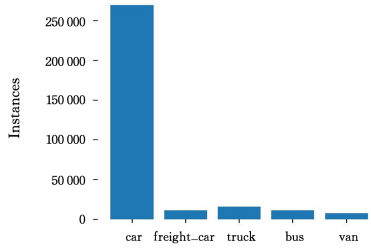
$$mAP=\frac{\sum_{i=0}^{N}AP_i}{N} \tag{14}$$

$$FPS=\frac{Frames}{Time} \tag{15}$$

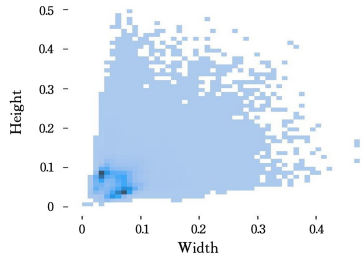
其中,TP(True Positives)指正确检测框的数量,FP(False Positives)指错误检测框的数量,FN(False Negatives)指漏检目标框的数量,AP(Average-Precision)为单个类别的目标检测精度,N 为所有类别的数量,$Frames$ 为视频帧数或图像数量,$Time$ 为检测所需时间。

4.3 数据集及其预处理

本文采用天津大学发布的大规模可见光/红外双模态图像数据集 DroneVehicle,该数据集由无人机搭载红外与可见光摄像头同步拍摄得到。拍摄完成后,天津大学团队将成对的可见光和红外图像进行了裁剪、缩放、padding 填充及匹配对齐处理。对齐处理后,可见光和红外图像分辨率均为 840×712 像素。DroneVehicle 共有 28439 对图像,其中训练集 17990 对,验证集 1469 对,测试集 8980 对。目标类别包括 car,freight car,truck,bus,van。拍摄时间包括白天(Day)、晚上(Night)、黑夜(Dark night)等不同场景,其中白天 14478 对,晚上 8493 对,黑夜 5468 对。图 9(a)为各类别目标的数量分布情况,图 9(b)为目标相对于图像宽高的比例情况。可以看出,各类别目标的数量分布不均衡,car 的数量远高于其他类别;大多数目标尺寸较小,与图像的宽高比例小于 0.1。



(a)各类别目标的数量情况



(b)标签的长宽情况

图 9 目标数量及标签长宽分布情况

Fig. 9 Number of objects and distribution of label sizes

原数据集将成对的红外图像和可见光图像分别进行了手工标注。由于可见光和红外图像是对齐的,且红外图像的标注质量更高,因此,我们采用红外图像的标注文件作为对应的可见光图像的标注文件。同时,为了更好地适应无人机航拍条件下弱小目标检测的需求,本文在保持原图像宽高比例不变的情况下,将图像缩小为 512×433 像素。

5 实验结果与分析

5.1 双模态图像融合前后的对比实验

本文以 YOLOv5s 模型为基准,分别测试了其利用可见光(RGB)、红外(IR)单模态图像,以及可见光与红外双模态图像拼接融合(RGB+IR)的检测结果,如表 1 所列。由于 DroneVehicle 数据集约一半数量的图像是在弱光及黑暗条件下拍摄的,而可见光图像中的目标特征较弱,甚至完全不可见,导致仅利用可见光图像检测时,检测精度明显低于红外图像。采用可见光与红外图像拼接融合后,检测精度比可见光图像提升了 12.78%,比红外图像提升了 1.31%,而模型参数量仅增加 $0.001\ 2\times10^6$,基本可以忽略不计。可见,采用通道拼接融合的方法是非常有效的,仅需要增加极少的参数量,即可有效利用可见光与红外图像的互补信息。

表 1 采用不同模态图像检测的结果对比

Table 1 Comparison of detection results using different modal images

模型	模态	car	freight car	truck	bus	van	P/%	R/%	mAP_{50}/%	参数量
YOLOv5s	RGB	88.65	45.87	55.72	88.61	41.21	68.60	59.98	64.01	7.0236×10^6
	IR	97.46	62.75	70.82	93.44	52.92	**77.18**	69.67	75.48	7.0236×10^6
	RGB+IR	**97.70**	**64.40**	**73.88**	**94.74**	**53.20**	76.34	**72.39**	**76.79**	7.0248×10^6

5.2 重参数化瓶颈模块消融实验

本节以 YOLOv5s 为基准算法，以 RGB+IR 双模态图像为输入，分别用 3.1 节中提出的 9 种重参数化瓶颈模块替换 YOLOv5s 骨干网中的瓶颈模块，检测结果如表 2 所列。

表 2 重参数化瓶颈模块的性能对比

Table 2 Performance comparison of re-parameterized bottleneck modules

(%)

RepBottleneck	P	R	mAP_{50}
基准算法	76.34	72.39	76.79
(a)	78.47	70.69	76.99
(b)	77.67	71.93	77.43
(c)	77.04	73.13	77.58
(d)	77.48	72.17	77.04
(e)	78.62	71.45	77.42
(f)	78.23	72.06	77.55
(g)	77.17	72.46	77.34
(h)	77.32	**73.33**	**77.69**
(i)	**79.17**	71.4	77.60

从表 2 可以看出，采用重参数化模块后，模型性能均得到了提升，但各模块带来的提升幅度并不相同。其中，RepBottleneck(a)，RepBottleneck(d)和 RepBottleneck(g)检测精度提升幅度较小，分别提升 0.2%，0.25%和 0.55%。这是因为在带有残差连接的结构中采用串联的重参数化模块，会出现串联的恒等映射分支或串联的 1×1 卷积分支，从而破坏特征图的多样性。

为了避免出现串联的恒等映射分支或串联的 1×1 卷积分支，本文在瓶颈结构中仅采用 1 个重参数化模块，即得到 RepBottleneck(b)(c)，RepBottleneck(e)，RepBottleneck(f)，RepBottleneck(h)和 RepBottleneck(i)。可以看出，采用本文所提 RepBlock_31_BN 模块的 RepBottleneck(h)和 RepBottleneck(i)性能更优，尤其是 RepBottleneck(h)检测精度相对于 YOLOv5s 基准提升了 0.90%。因此，本文在骨干网中最终采用 RepBottleneck(h)模块。

5.3 多层次特征融合模块消融实验

对于特征金字塔的最浅层 P3 层，本文将 MLFFM-P3 模块的卷积核尺寸 K 分别设置为[3,5,7,9]，[3,3,5,7]，[1,3,5,7]进行消融实验。为了减少参数量和计算量，对 5×5，7×7，9×9 大尺寸卷积核采用分组卷积策略，分组数分别设置为 2，4，8，实验结果如表 3 所列。可见，在 P3 层中采用过大或过小的卷积核尺寸均无法取得最佳结果，当 K=[3,3,5,7]，G=[1,1,2,4]时性能最好，检测精度达到 77.50%。实际上，如果不采用分组卷积，本文所提多层次特征融合模块的性能可进一步提升，但会增加一定参数量与计算量，因此本文并未采用。

表 3 MLFAM-P3 模块消融实验

Table 3 Ablation experiments of MLFAM-P3 module

K	G	mAP_{50}/%
[3,5,7,9]	[1,2,4,8]	77.06
[3,3,5,7]	[1,1,2,4]	**77.50**
[1,3,5,7]	[1,1,1,2]	77.26

针对特征金字塔的 P4 层，由于其有效感受野相较于 P3 层进一步缩小，因此，将 MLFFM-P4 模块的卷积核尺寸也适当缩小，将卷积核 K 分别设置为[3,3,5,7]，[1,3,5,7]，[1,3,3,5]，将 5×5 和 7×7 卷积的分组数分别设为 2 和 4，实验结果如表 4 所列。可见，当 K=[1,3,3,5]，G=[1,1,1,2]时性能最好。

表 4 MLFAM-P4 模块消融实验

Table 4 Ablation experiments of MLFAM-P4 module

K	G	mAP_{50}/%
[3,3,5,7]	[1,1,2,4]	76.91
[1,3,5,7]	[1,1,2,4]	77.00
[1,3,3,5]	[1,1,1,2]	**77.06**

5.4 不同改进对模型性能的影响

为了验证本文所提改进方法的有效性，本文以 YOLOv5s 模型作为基准，将可见光和红外双模态图像拼接融合后作为输入，逐一测试了各改进方法对模型性能的影响，如表 5 所列。

表 5 不同改进对模型性能的影响

Table 5 Effect of different improvements on model performance

模型	RepCSPDarkNet	MLFFM-P3	MLFFM-P4	Without P5	P/%	R/%	mAP_{50}/%	参数量
YOLOv5s	−	−	−	−	76.34	72.39	76.79	7.0248×10^6
	√	−	−	−	77.32	**73.33**	77.69	8.0160×10^6
	−	√	−	−	76.71	72.88	77.50	7.0646×10^6
	−	−	√	−	76.77	72.20	77.06	7.1377×10^6
	√	√	√	−	77.76	72.88	77.92	8.1696×10^6
	√	√	√	√	**78.39**	72.46	**77.93**	$\mathbf{6.3829\times10^6}$

YOLOv5s 骨干网采用 RepCSPDarkNet 后，检测精度 mAP_{50} 达到 77.69%，较原模型提升了 0.90%，骨干网的特征提取能力明显增强。在 P3 特征层中采用 MLFFM-P3 模块后，mAP_{50} 提升了 0.71%。在 P4 特征层中采用 MLFFM-P4 模块后，mAP_{50} 提升了 0.27%。同时采用 RepCSPDarkNet，MLFFM-P3 和 MLFFM-P4，检测精度 mAP_{50} 达到 77.92%，相比原模型提升了 1.13%。在此基础上进一步删除 P5 检测层后，参数量减少了 1.79×10^6，而检测精度 mAP_{50} 不仅没有

降低，反而提升了 0.01%。可见，P5 检测层对于弱小目标检测而言是冗余的，不利于弱小目标的检测。

5.5 与主流算法的对比实验

在 DroneVehicle 数据集上，本文提出的 DM-YOLO 模型与基准算法 YOLOv5s，以及目前先进的实时目标检测算法 YOLOv6 和 YOLOv7 进行了对比。由于 YOLOv6 与 YOLOv7 有多个不同规模的模型，因此本文选取了规模相近的模型 YOLOv6 n/s 以及 YOLOv7-Tiny 进行比较，实验结果如表 6 所列。

当仅利用可见光(RGB)或红外(IR)单模态图像检测时，与基准算法 YOLOv5s 相比，本文提出的 DM-YOLO 在参数量更少的情况下，对可见光图像的检测精度高出 2.1%，对红外图像的检测精度高出 0.56%；与 YOLOv6s 相比，在参数量仅为其 34.50%、计算量为其 40.47%的情况下，对可见光图像的检测精度高出 1.31%，对红外图像的检测精度与其相差不大；与 YOLOv7-Tiny 相比，虽然参数量、计算量略有增加，但本文方法对可见光图像和红外图像的检测精度分别高出 12.91%和 11.24%。

表 6 与目前主流的实时目标检测算法比较

Table 6 Comparison with current mainstream real-time object detection algorithms

模型	模态	car	freight car	truck	bus	van	P/%	R/%	mAP_{50}/%	参数量	*FLOPs*	*FPS*
YOLOv5s	RGB	88.65	45.87	55.72	88.61	41.21	68.60	59.98	64.01	7.0236×10^6	10.1×10^9	109
	IR	97.46	62.75	70.82	93.44	52.92	77.18	69.67	75.48			
YOLOv6n	RGB	84.50	28.30	47.10	83.70	30.90	60.10	55.00	54.90	4.6300×10^6	7.26×10^9	121
	IR	97.60	54.00	68.20	92.40	41.00	69.30	68.60	70.60			
YOLOv6s	RGB	90.20	45.30	58.00	89.40	41.30	68.50	62.40	64.80	18.5000×10^6	28.91×10^9	76
	IR	**98.20**	62.80	**75.60**	93.80	51.10	75.50	72.40	76.30			
YOLOv7-Tiny	RGB	86.50	29.80	43.00	83.20	23.60	56.50	52.30	53.20	6.0184×10^6	8.40×10^9	118
	IR	97.00	45.80	58.90	90.90	31.40	63.60	64.40	64.80			
DM-YOLO	RGB	89.95	48.26	58.27	89.65	44.43	69.42	62.47	66.11	6.3817×10^6	11.60×10^9	83
	IR	97.63	62.33	72.54	93.74	53.94	76.02	72.17	76.04			
	RGB+IR	97.91	**64.34**	74.63	**95.05**	**57.74**	**78.39**	**72.46**	**77.93**	6.3829×10^6	11.70×10^9	82

当本文 DM-YOLO 模型利用双模态图像融合检测时，检测精度进一步提升，分别比 YOLOv5s，YOLOv6n，YOLOv6s，YOLOv7-Tiny 的最高精度高出 2.45%，7.33%，1.63%，13.13%，取得了明显优于其他模型的检测精度。可见，YOLOv6 和 YOLOv7 模型虽然在通用目标检测数据集 MS COCO 上取得了优异的性能，但对于复杂场景下的弱小目标检测效果并不理想。而本文 DM-YOLO 模型对弱小目标具有更强的特征提取能力，能更好地适应复杂天候条件下的目标检测任务。当利用可见光与红外图像融合检测时，DM-YOLO 模型的检测速度达到 82 FPS，虽相较于 YOLOv5s 和 YOLOv7-Tiny 略有差距，但完全可以满足实时检测的任务需求。DM-YOLO 检测速度略慢的原因主要是其引入注意力机制后，增加了神经网络层数，导致推理时间增加。

5.6 检测效果及分析

为了更直观地观察检测效果，本文选取了白天、晚上和黑夜等不同天候条件下的图像进行检测，效果如图 10 所示。第一列为标注图像，YOLOv6s(RGB)和 YOLOv6s(IR)分别为 YOLOv6s 对可见光图像、红外图像的检测结果；YOLOv7 Tiny(RGB)和 YOLOv7 Tiny(IR)分别为 YOLOv7 Tiny 对可见光图像、红外图像的检测结果；最后一列为本文提出的 DM-YOLO 的检测结果。其中，红色虚线框表示漏检目标，黄色虚线框表示误检目标。

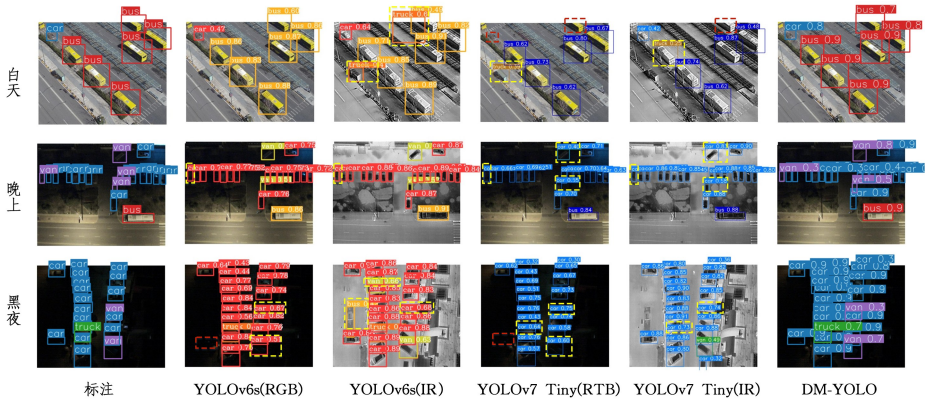


图 10 检测效果对比图(电子版为彩图)

Fig. 10 Comparison of detection effects

在第一行白天光照良好的条件下，YOLOv6s(RGB)对图中目标全部检测正确，效果相对较好；YOLOv6s(IR)则有两例误检；YOLOv7 Tiny(RGB)对树木遮挡的 car 和残缺的 bus 漏检，并有 1 例误检；YOLOv7 Tiny(IR)则有 1 例误检，1 例漏检；本文 DM-YOLO 则全部检测正确，且置信度值总体高于其他模型。在第二行晚上弱光条件下，YOLOv6s(IR)和 YOLOv6s(RGB)各有 1 例误检，但 YOLOv6s(IR)的目标置信度值整体高于 YOLOv6s(RGB)；YOLOv7 Tiny(IR)的目标置信度值也整体高于 YOLOv7 Tiny(RGB)；DM-YOLO 由于能够利用可见光和红外图像的互补信息，对所有目标全部

检测正确。在第三行黑夜条件下，由于光线非常微弱，人眼在可见光图像中基本无法发现目标。YOLOv6s(RGB)利用微弱的光线虽能检测出大部分目标，但目标置信度普遍偏低，且出现1例漏检、2例误检；YOLOv6s(IR)虽能检测出全部目标且目标置信度相对较高，但由于红外目标的特征信息有限，将阴影误检为了bus，将特征相似的van误检为了car；YOLOv7 tiny检测效果与YOLOv6s类似；而本文DM-YOLO对所有目标均能正确检测，且目标置信度更高。

总体来看，在光照条件良好的条件下，利用可见光图像检测效果较好，在黑暗条件下，利用红外图像检测效果较好，而DM-YOLO能够综合利用双模态图像的互补信息，对弱小目标有更强的特征提取能力，在不同天候条件下均表现较好。

结束语 针对复杂天候条件下无人机航拍目标检测存在的问题，本文提出了重参数化方法增强的双模态实时目标检测模型DM-YOLO。采用通道拼接的方法对双模态图像进行融合，以极低的融合成本有效利用了双模态图像的互补信息。本文提出了更加高效的重参数化模块，并以此为基础构建了更强大的骨干网RepCSPDarkNet，有效增强了骨干网对弱小目标的特征提取能力。同时，提出了多层次特征融合模块，可更充分提取并聚焦弱小目标的多尺度特征。最后，通过删除冗余的深层检测层，减少了模型参数量。与目前先进的实时目标检测算法YOLOv6以及YOLOv7相比，本文方法取得了良好的精度与速度的平衡，未来可应用于无人机终端在复杂天候条件下执行目标检测任务，具有良好的应用前景。

由于DroneVehicle数据集存在长尾分布现象，freight car，truck，van等类别的样本数量明显少于car的数量，导致其检测精度明显低于car。对于长尾分布问题，本文并未进行处理，未来将继续研究改进目标损失函数，通过增加稀有类别的损失比重，提高稀有类别的检测精度。此外，实验表明对freight car，truck，van等的检测精度不仅低于car，还低于样本数量大致相同的bus。这可能是由于它们外形相似、特征差异小，因此容易出现误检。对此，未来将借鉴细粒度识别领域的研究工作，以提高模型对相似目标的细粒度识别能力。

参 考 文 献

[1] NIU W H，YIN M M. Road Small Target Detection Algorithm Based on Improved YOLOv5[J]. Chinese Journal of Sensors and Actuators，2023，36(1)：36-44.

[2] XIE P X，CUI J R，ZHAO M. Electiric Bike Helment Wearing Detection Alogrithm Based on Improved YOLOv5[J]. Computer Science，2023，50(S1)：420-425.

[3] YANG Y H，ZHONG B J，TIAN H W. Target Detection Model of DS-yolov4-Tiny Rescue Robot[J]. Computer Simulation，2022，39(1)：387-393.

[4] GIRSHICK R B，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. Columbus：IEEE，2014：580-587.

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

[6] REN S Q，HE K M，GIRSHICK R B，et al. Faster R-CNN：Towards Real-Time Object Detection with Region Proposal Networks[C]// Conference and Workshop on Neural Information Processing Systems. Montreal：MIT Press. 2015：91-99.

[7] LIN T Y，DOLLÁR P，GIRSHICK R B，et al. Feature Pyramid Networks for Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：936-944.

[8] CAI Z W，VASCONCELOS N. Cascade R-CNN：Delving Into High Quality Object Detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City：IEEE，2018：6154-6162.

[9] 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. Las Vegas：IEEE，2016：779-788.

[10] REDMON J，FARHADI A. YOLO9000：Better，Faster，Stronger[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Honolulu：IEEE，2017：6517-6525.

[11] REDMON J，FARHADI A. YOLOv3：An Incremental Improvement[J]. arXiv：1804.02767，2018.

[12] BOCHKOVSKIY A，WANG C Y，LIAO H Y. YOLOv4：Optimal Speed and Accuracy of Object Detection[J]. arXiv：2004.10934，2020.

[13] ULTRALYTICS. YOLOv5 [EB/OL]. https：//github.com/ultralytics/yolov5.

[14] LI C，LI L L，JIANG H L，et al. YOLOv6：A Single-Stage Object Detection Framework for Industrial Applications[J]. arXiv：2209.02976，2022.

[15] WANG C Y，ALEXEY B，MARK L，et al. YOLOv7：Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023：7464-7475.

[16] LIU W，ANGUELOV D，ERHAN D，et al. SSD：Single Shot MultiBox Detector[C]// Proceedings of the European Conference on Computer Vision. Amsterdam：Springer. 2016：21-37.

[17] FU C Y，LIU W，RANGA A，et al. DSSD：Deconvolutional Single Shot Detector[J]. arXiv：1701.06659，2017.

[18] WU Z，MIAO X D，LI W W，et al. Low-Visibility Road Target Detection Algorithm Based on Infrared and Visible Light Fusion[J]. Infrared Technology，2022，44(11)：1154-1160.

[19] 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 Conference on Computer Vision and Pattern Recognition，2022：5792-5801.

[20] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]// Proceedings of the International Conference on Neural Information Processing Systems. Montreal，2014：2672-2680.

[21] GENG K K，ZOU W，YIN G D，et al. Low-observable targets detection for autonomous vehicles based on dual-modal sensor fusion with deep learning approach[J]. Journal of Automobile

Engineering, 2019, 233(9): 2270-2283.
[22] ZHOU H, SUN M, REN X, et al. Visible-Thermal Image Object Detection via the Combination of Illumination Conditions and Temperature Information[J]. Remote Sensing, 2021, 13(18): 36-56.
[23] CHEN Y T, SHI J G, YE Z L, et al. Multimodal Object Detection via Probabilistic Ensembling[C]// Proceedings of the European Conference on Computer Vision. 2022(9): 139-158.
[24] 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: 6700-6713.
[25] HE K M, ZHANG X Y, REN S Q, et al. Deep Residual Learning for Image Recognition[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
[26] DING X H, ZHANG X Y, MA N N, et al. RepVGG: Making VGG-Style ConvNets Great Again[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 13733-13742.
[27] DING X H, ZHANG X Y, HAN J G, et al. Diverse Branch Block: Building a Convolution as an Inception-Like Unit[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 10886-10895.
[28] KUMAR P, GABRIEL J, ZHU J, et al. MobileOne: An Improved One millisecond Mobile Backbone[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2023: 7907-7917.
[29] SANDLER M, HOWARD A, ZHU M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510-4520.
[30] HOWARD A, SANDLER M, CHU G, et al. Searching for MobileNetV3[C]// Proceedings of the 2019 IEEE International Conference on Computer Vision. 2019: 1314-1324.
[31] MA N N, ZHANG X Y, ZHENG H T, et al. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design[C]// Proceedings of the European Conference on Computer Vision. 2018(14): 122-138.
[32] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More Features From Cheap Operations[C]// Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition. 2020: 1577-1586.
[33] HAN K, WANG Y H, XU C, et al. GhostNets on Heterogeneous Devices via Cheap Operations[J]. International Journal of Computer Vision, 2022, 130: 1050-1069.
[34] CHEN C P, GUO Z C, ZENG H E, et al. RepGhost: A Hardware-Efficient Ghost Module via Re-parameterization[J]. arXiv: 2211.06088, 2022.
[35] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE. 2018: 7132-7141.
[36] WOO S, PARK J, LEE J Y, et al. CBAM: convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision. Munich: Springer. 2018, 11211: 3-19.
[37] HOU Q B, ZHOU D Q, FENG J S. Coordinate Attention for Efficient Mobile Network Design[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Virtual: IEEE. 2021: 13713-13722.
[38] ZHANG H, ZU K K, LU J, et al. EPSANet: An Efficient Pyramid Squeeze Attention Block on Convolutional Neural Network[C]// Proceedings of the Asian Conference on Computer Vision. 2022: 541-557.

LI Yunchen, born in 1987, postgraduate. His main research interest is object detection.

ZHANG Rui, born in 1977, Ph.D, professor, Ph.D supervisor. His main research interests include data engineering and information fusion.

（责任编辑：何杨）