◆生产场景◆

基于多尺度特征提取的密集型小目标检测网络*

元昌安[1,2],王文姬[1],黄豪杰[3],覃正优[1],张金勇[1],廖惠仙[4],覃　晓[1,5**],李小森[6],李永玉[1],
符云琴[1],谭思婧[1],钱泉梅[1],吴琨生[7]
(1.南宁师范大学,广西人机交互与智能决策重点实验室,广西南宁　530100;2.广西科学院,广西南宁　530007;3.中国通信服务股份有限公司广西技术服务分公司,广西南宁　530000;4.广东财贸职业学院数字技术学院,广东清远　511510;5.广西区域多源数据集成与智能处理协同创新中心,广西桂林　541004;6.广西民族大学人工智能学院,广西南宁　530006;7.广西壮族自治区南宁树木园,广西南宁　530225)

摘要:针对现有的无锚框目标检测算法难以在密集场景下有效提取多尺度目标特征的问题,本研究提出基于多尺度特征提取的密集型小目标检测网络(Intensive small target detection network based on Multi-Scale feature Extraction,IMSE)。本研究首先提出多尺度特征增强(Multi-scale Feature Enhancement,MFE)模块,其包括窗口注意力(Window Attention,WA)模块和多尺度信息融合(Multi-scale Information Fusion,MIF)模块,通过建立全局级别的上下文联系从而增强 IMSE 在密集场景下的特征表达,进而能够更有效地提取检测目标的多尺度特征;其次提出可变形卷积特征金字塔网络(Deformable Convolutional Feature Pyramid Networks,DCFPN)结构,引入空洞卷积进行特征增强,从而能够有效提高 IMSE 检测形状不规则、分布无规律物体的能力;最后将融合后的多尺度特征分别输入检测头进行分类与边界框的回归任务。IMSE 在公共数据集 MS COCO、CARPK 与基于实际生产场景构建的 WOOD 数据集上进行验证,实验结果表明,IMSE 在 3 个数据集上的平均精度(Average Precision,AP)分别达到了 49.4%、75.8%和 55.0%,分别比原始 FCOS 方法高出 1.8%、1.4%和 2.1%,验证了所提出模型的有效性。

关键词:目标检测;自注意力机制;特征金字塔;空洞卷积;可变形卷积

中图分类号:S781,TP391.41,TP183　　**文献标识码**:A　　**文章编号**:1005-9164(2024)05-0939-15
DOI:10.13656/j.cnki.gxkx.20241127.011

密集型小目标检测是指在图像中识别和定位大量小而密集分布的目标对象。随着深度学习技术的发展,密集型小目标检测在计算机视觉领域中具有广泛的应用价值,例如人群计数、交通流量分析、农作物

收稿日期:2024-07-22　　修回日期:2024-09-24
*广西科技重大专项(桂科 AA22068057 和桂科 AB21076021)资助。
【第一作者简介】
元昌安(1964—),男,博士,教授,主要从事智能计算研究,E-mail:yuanchangan@126.com。
【**通信作者简介】
覃　晓(1973—),女,硕士,教授,主要从事数字图像处理、自然语言理解研究,E-mail:7670172@qq.com。
【引用本文】
元昌安,王文姬,黄豪杰,等.基于多尺度特征提取的密集型小目标检测网络[J].广西科学,2024,31(5):939-953.
YUAN C A,WANG W J,HUANG H J,et al. An Intensive Small Object Detection Network Based on Multi-scale Feature Extraction [J]. Guangxi Sciences,2024,31(5):939-953.

检测等[1-3]。与传统的目标检测任务不同,密集型小目标检测任务中的待检测物体具有数量多、形状不规则、尺寸不统一、分布密集、像素数量较少等特点,这都对现有的检测技术如何准确地识别和定位图像中的多个小目标提出新的挑战。

从生成目标框的角度可以将现有的密集型小目标检测算法分为基于锚框(Anchor-based)检测和无锚框(Anchor-free)检测两类。基于锚框的检测模型通过产生密集的锚框以生成高质量的候选区域,通常具有较高的准确性。如杨攀等[4]提出一种基于 Mask R-CNN(Mask Region-based Convolutional Neural Network)的木材分割方法,能适应各种场景下的各尺寸大小密集木材检测分割任务。霍爱清等[5]通过改进 YOLOv3(You Only Look Once version 3)算法,解决车辆密集场景中检测目标重叠率高而导致的漏检、误检问题。基于锚框的目标检测模型需要人工设定锚框和依靠 RPN[6](Region Proposal Network)网络生成候选框,导致模型更依赖于先验知识而缺乏泛化能力。基于无锚框的目标检测方法不依赖锚框就能直接预测出边界框和目标类别,有效解决了由于预设固定尺度的锚框导致模型在检测尺寸或纵横比显著差异的物体时性能不佳的问题。因此,基于无锚框的目标检测方法能够更好地完成密集场景下的目标检测任务。如 Wu 等[7]为了使用无人机检测密集的杏花,提出了密集目标检测方法 D-YOLOv8,且改进后的网络可以有效地支持任何密集目标检测任务。此外,基于 Transformer[8]的目标检测模型 DETR[9],在没有预定义先验锚框和 NMS(Non-Maximum Suppression)后处理策略的情况下能够实现端到端的目标检测。但是基于 Transformer 的检测模型计算成本较高,硬件资源消耗较大,一般的计算环境难以支撑其庞大的计算需求。综上,本研究选用基于 CNN 的无锚框检测算法作为研究基础。

FCOS[10]是基于无锚框的全卷积目标检测方法之一,其通过判断图像中每一个像素点的目标属性及其边界框的存在以完成目标定位任务,相比于其他基于无锚框的方法,FCOS 无需频繁构建锚定点方案,节省训练所需的内存,提高速度和检测性能,能有效地处理各种大小和形状不同的目标。具体而言,首先 FCOS 网络在骨干网络(Backbone)以卷积的计算方式对输入图像进行特征提取,在不同阶段得到不同尺度的特征图;其次在特征金字塔网络(Feature Pyramid Network,FPN)[11]中将提取到的高级语义特征上采样后与低级特征进行融合;最后分别对融合得到的不同尺度的特征在检测头(Head)部分进行逐像素的目标分类与边界框回归。Nayak 等[12]在无人机检测任务中验证了 FCOS 识别小目标的性能;Fu 等[13]在苎麻植株数量检测中发现 FCOS 与其他模型相比更适合在复杂情况下识别小物体作物。

尽管基于 FCOS 的改进工作已经取得了不错的进展,但是仍存在以下两个问题:①FCOS 中用于特征提取的 Backbone 均由传统卷积层下采样堆叠而成。由于卷积是固定感受野下的局部操作,难以高效地挖掘图像的全局上下文特征,因此现有的解决方案是在 Backbone 中引入注意力机制[14-16]。但是这些注意力机制仅局限于对通道、空间、局部或全局等单个方面进行特征关注,而缺少任何一方面的信息都有可能导致网络的精度不高和泛化性不足。同时,传统自注意力机制的引入也容易使模型过于关注全局信息而忽略局部信息,缺乏对语义信息的提取,且计算成本的增加也是不可忽视的问题。②FCOS 特征融合模块由传统的 FPN 架构组成,现有大部分针对基于 FPN 的改进工作仍保留着传统卷积的融合方法[17-20]。传统卷积是利用固定长宽比的滑动窗口在特征图上进行滑动计算,然而这种计算方式在对密集且不规则的物体进行检测时容易丢失描述目标信息的像素点,忽略可能仅存在于特定级别的关键细节,导致信息丢失和特征提取不完整,使得传统 FPN 不能很好地适应密集场景下的特征融合。

综上所述,针对传统 FCOS 难以在复杂背景中提取到丰富的全局上下文特征,以及在密集、形状不一、尺寸不一致的物体检测场景下容易出现漏检、错检和重检的问题,本研究提出了基于多尺度特征提取的密集型小目标检测网络(Intensive small target detection network based on Multi-Scale feature Extraction,IMSE)。IMSE 主要包含两个模块:多尺度特征增强(Multi-scale Feature Enhancement,MFE)模块和可变形卷积[21]特征金字塔网络(Deformable Convolutional Feature Pyramid Network,DCFPN)结构,其中 MFE 模块包括窗口注意力(Window Attention,WA)模块和多尺度信息融合(Multi-scale Information Fusion,MIF)模块两部分。首先,WA 模块在特征提取阶段加入了多尺度注意力机制,引导网络更全面地学习图像的特征信息;然后,MIF 模块通过将输入特征图并行经过一系列卷积和池化操作,

指导网络在更多维度上融合图像特征,捕捉特征元素的全局和上下文信息;最后,通过 DCFPN 结构使网络在特征融合阶段能够更好地融合 MFE 模块的输出特征,提高模型对形状不规则、分布无规律的物体特征的表征能力。

1 模型与方法

本研究提出的 IMSE 总体架构如图 1 所示。IMSE 由 3 部分组成:骨干网络、DCFPN 结构和检测头。首先,IMSE 使用 ResNet-101 作为骨干网络进行特征提取,与 FCOS 网络结构相比,IMSE 在骨干网络的输出层 C3、C4 和 C5 后添加 MFE 模块,将骨干中生成的不同分辨率的特征图分别输入 WA 模块和 MIF 模块进行特征提取,引导网络更全面地关注全局上下文信息,提高网络在特征提取阶段的多尺度表达能力;然后,将 WA 模块和 MIF 模块得到的特征进行相加融合,并将融合后的特征输入 DCFPN 结构,DCFPN 结构引入可变形卷积来融合语义和不同尺度的特征,从而保留更多的语义信息,生成多尺度特征图作为输出;最后,DCFPN 结构输出的特征和传统 FPN 结构的两个标准卷积层被分别输入到每一个检测头中,进行分类与边界框的回归任务。

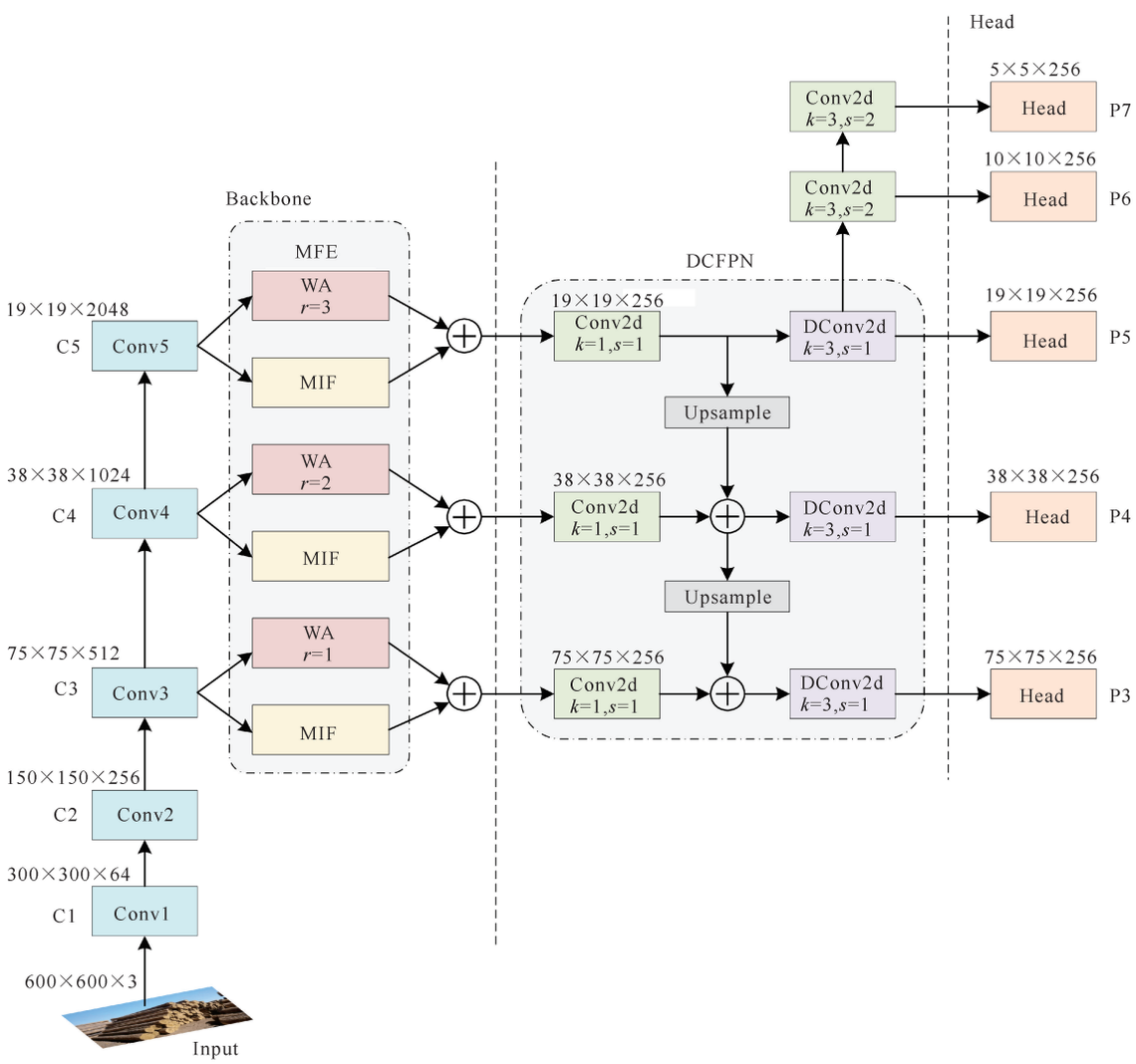


图 1 IMSE 结构

Fig. 1 Structure of IMSE

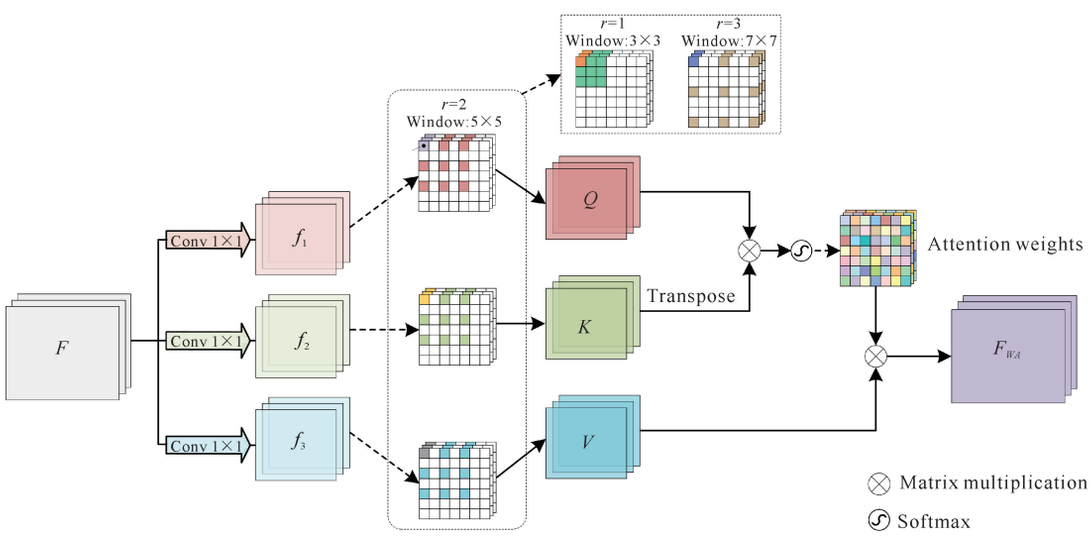
1.1 MFE 模块

MFE 模块由 WA 模块和 MIF 模块构成。WA 模块利用自注意力机制在不同尺度上提取空间特征,使模型更好地理解全局上下文的语义信息,捕捉到长距离的语义依赖关系,理解不同尺度的特征之间的相互关系,提高模型的表达能力。MIF 模块通过不同

大小的卷积和全局池化操作,实现局部空间的多尺度特征和通道特征的提取及深度融合,进一步挖掘图像的轮廓、纹理、颜色和边缘等更加丰富的细节信息。最终,相加融合 WA 模块和 MIF 模块提取的特征能够在全局和局部、空间和通道以及语义和定位上全面地学习和理解图像特征,达到特征增强的效果。

1.1.1 WA 模块

为了在含有遮挡或较多干扰噪声的复杂环境下有效地检测小目标,本研究设计 WA 模块以加强全局上下文联系,提高网络对目标的精确定位和识别能力。WA 模块能够在传统的目标检测特征提取阶段,通过引入多尺度自注意力机制增强特征。但自注意力机制的 3 个参数 Q、K、V 的计算需要消耗较大的计算资源。为了降低计算代价,本研究设计分段计算 Q、K、V 的算法 Seg_QKV。Seg_QKV 利用大小为 $\omega\times\omega$ 的窗口遍历输入的特征图,在遍历过程中每读取一个窗口的特征子图,就计算对应的权重矩阵(Q_ω、K_ω、V_ω)。遍历完成后,将所有的特征子图进行拼接,得到整张特征图的自注意力参数(Q,K,V)。同时,为了更好地关注多尺度特征,本研究在设计窗口时引入了空洞卷积[22]的思想,通过设置不同的空洞系数,得到不同大小的窗口,获得不同尺度的注意力权重。WA 模块的结构如图 2 所示。



The input feature F corresponds to the $C_{[i]}$-layer output feature map in the backbone network, where $i\in 3,4,5$, and F_{WA} represent the output of the module.

图 2 WA 模块结构

Fig. 2 Structure of WA module

以滑动窗口大小 5×5 为例,WA 模块分两步执行,第一步是窗口注意力参数 Q、K、V 计算,第二步是进行多尺度特征提取。本研究使用两个算法来实现 WA 模块,一是分段计算 Q、K、V 的算法 Seg_QKV,二是多尺度注意力计算算法 MSa。

①分段计算 Q、K、V 算法 Seg_QKV

该部分借鉴了卷积操作中的滑动窗口思想,利用滑动窗口的形式对特征图进行提取,将自注意力的计算限制在滑动窗口内进行计算,从而达到快速计算的效果,具体流程如算法 1 所示。

算法 1 Seg_QKV

输入:特征图 $F=[H,W,C]$,滑动窗口大小 $\omega\times\omega$,滑动步长 S

输出:权重矩阵 Q_ω,K_ω,V_ω

1: $Q_\omega=[]$,$K_\omega=[]$,$V_\omega=[]$;
2: for $i=1$ to 3 do
3: $f_i=\mathrm{Conv}(1\times1,F)$; //表示对 F 进行 1×1 卷积操作
4: end for
5: for $i=1$ to C do
6: for $a=1$ to $(W-\omega)$ do // a 表示滑动窗口在特征图 F 上左上角横坐标
7: for $b=1$ to $(H-\omega)$ do // b 表示滑动窗口在特征图 F 上左上角纵坐标
8: $\mathrm{Win}_{ab}^{Q}=f_1^{ab}[K,K,1]$; // Win_{ab}^{Q} 表示左上角坐标为 (a,b) 的 $K\times K$ 窗口在特征

图 f_1 上所提取到的特征图，其特征维度为 $[K,K,1]$

$Q_\omega=\text{Concat}(Q_\omega,\text{Win}_{ab}^{Q})$；

$\text{Win}_{ab}^{K}=f_2^{ab}[K,K,1]$；

$K_\omega=\text{Concat}(K_\omega,\text{Win}_{ab}^{K})$；

$\text{Win}_{ab}^{V}=f_3^{ab}[K,K,1]$；

$V_\omega=\text{Concat}(V_\omega,\text{Win}_{ab}^{V})$；

$b=b+S$；

end for

$a=a+S$；

end for

end for

return $Q_\omega,K_\omega,V_\omega$

②多尺度注意力计算算法 MSa

空洞卷积是在增加感受野的同时保持特征图尺寸不变的操作，弥补了特征图在缩小和放大过程中造成的信息损失，同时通过改变扩张率可以增加感受野。通过不同的空洞系数可得到不同大小的滑动窗口，然后对特征图进行特征提取，从而达到多尺度的效果，具体流程如算法 2 所示，该算法名为 MSa (Multi-Scale attention)。考虑特征提取的局部性与全局性，所采用的空洞系数分别为 1、2、3，滑动步长均为 1。

算法 2 MSa

输入：特征图序列 $F=(F_1,F_2,\cdots,F_n)$，空洞系数 $D=(D_1,D_2,\cdots,D_n)$，滑动步长 S

输出：特征图序列 $F_{\text{WA}}=(F_{\text{WA}[1]},F_{\text{WA}[2]},\cdots,F_{\text{WA}[n]})$

for $i=1$ to n do　　//索引 i 对应空洞系数的取值，当前 n 取值 3

$Q_i,K_i,V_i=\text{Seg\_QKV}(F_i,\text{Win}[2D_i+1],S)$；　// $\text{Win}[2D_i+1]$ 表示滑动窗口的大小为 $2d_i+1$

$F_{\text{WA}[i]}=\text{softmax}(Q_i K_i^{\text{T}}/\sqrt{d_k})V_i$；　// d_k 表示 K_i 的维度

end for

return F_{WA}

③关于算法 SC 时间复杂度与传统 Q、K、V 计算时间复杂度比较的讨论

传统的自注意力计算是将输入特征通过 3 个不同的线性投影层进行映射，从而得到 Q、K、V 3 个特征矩阵，随后根据 Softmax 式进行注意力的计算。传统的自注意力计算公式如公式(1)所示：

$$\text{Attention}(Q,K,V)=\text{Softmax}(\frac{QK^{\text{T}}}{\sqrt{d_k}})V, \quad (1)$$

由式(1)可知，注意力的计算过程为矩阵相乘。假设 Q、K、V 3 个特征矩阵大小为 n 行 d 列，滑动窗口大小为 3×3，则传统自注意力的时间复杂度 O_1 与本研究提出的 WA 模块注意力的时间复杂度 O_2 的计算公式如下所示：

$$O_1=O_{QK^{\text{T}}}+O_{(QK^{\text{T}})V}, \quad (2)$$

$$O_{QK^{\text{T}}}=n\times d\times n=n^2d, \quad (3)$$

$$O_{(QK^{\text{T}})V}=n\times n\times d=n^2d, \quad (4)$$

其中，QK^{T} 与 $(QK^{\text{T}})V$ 分别表示自注意力计算中矩阵 Q、K 相乘和矩阵 Q、K 相乘的结果与矩阵 V 相乘。由公式(2)至(4)可知，传统自注意力的时间复杂度 $O_1=2n^2d$。

$$O_2=O_{Q_{3\times3}K_{3\times3}^{\text{T}}}+O_{(Q_{3\times3}K_{3\times3}^{\text{T}})V_{3\times3}}, \quad (5)$$

$$O_{Q_{3\times3}K_{3\times3}^{\text{T}}}=(d-2)(n-2)\times3\times3\times3=27\times(d-2)(n-2), \quad (6)$$

$$O_{(Q_{3\times3}K_{3\times3}^{\text{T}})V_{3\times3}}=(d-2)(n-2)\times3\times3\times3=27\times(d-2)(n-2), \quad (7)$$

其中，$Q_{3\times3}K_{3\times3}^{\text{T}}$ 表示滑动窗口大小为 3×3 时的矩阵 Q、K 相乘，$(Q_{3\times3}K_{3\times3}^{\text{T}})V_{3\times3}$ 表示滑动窗口大小为 3×3 时矩阵 Q、K 相乘的结果与矩阵 V 相乘。因为滑动窗口是以步长 1 在输入特征矩阵上进行滑动，因此 $(d-2)(n-2)$ 表示滑动窗口在输入特征矩阵上的滑动次数，也代表着滑动窗口在输入特征矩阵上滑动结束后所得到的输出特征矩阵个数。由公式(5)至(7)可知，WA 模块注意力的时间复杂度 $O_2=54\times(d-2)(n-2)$。经过对比可知，$O_2<O_1$，体现出 WA 模块的计算速度，解决了传统自注意力机制存在的计算复杂度高的问题。

1.1.2 MIF 模块

特征图的空间信息包含物体位置信息，通道信息包含物体颜色信息，有效挖掘这两种信息对提高密集场景下小目标的检测性能至关重要。为解决现有注意力机制对特征信息关注不全面的问题，本研究提出 MIF 模块，结构如图 3 所示。与 WA 模块利用多尺度注意力机制提取特征不同，MIF 模块由一系列卷积和池化操作构成。池化操作只改变特征图的空间维度而不改变通道维度，能较好地突出特征图的通道特征。同时，利用多尺度空洞卷积、平均池化和最大池化对输入特征图进行并行操作，能更好地提取、融

合特征图的空间与通道信息。拼接融合各部分的局部空间特征后,分别将其与经过归一化和激活操作后的通道特征进行矩阵相乘来获得权重矩阵,最后将权重矩阵相加融合,实现特征的深度融合。

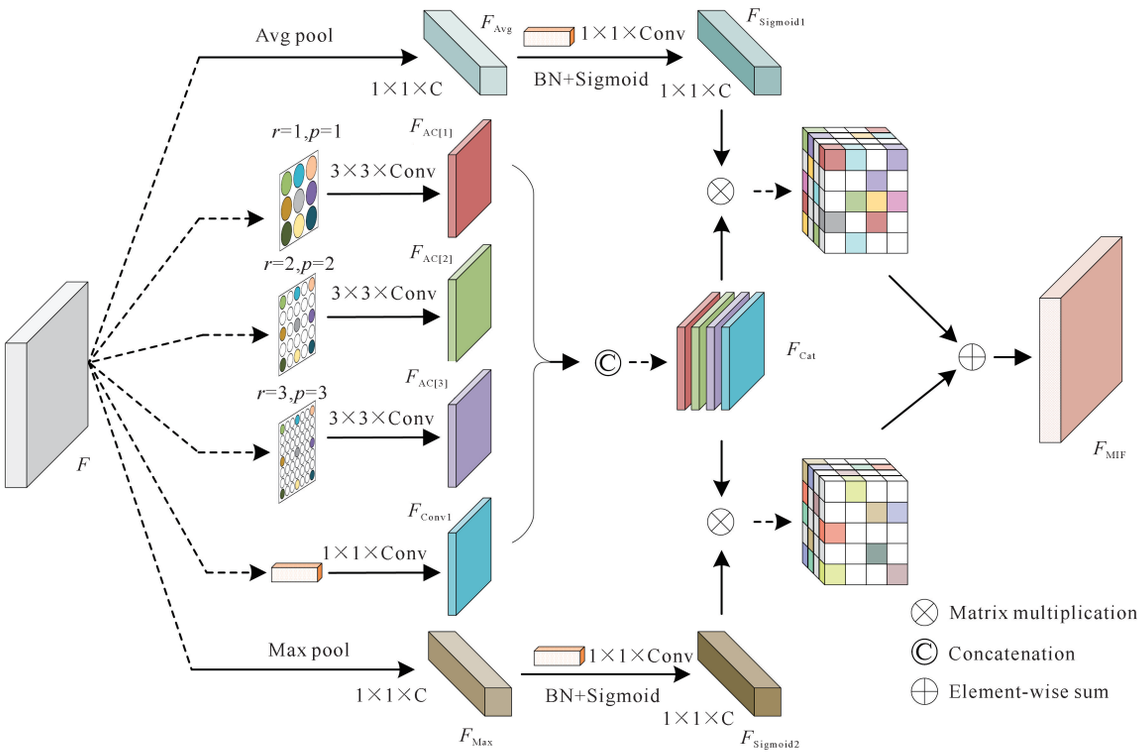


图 3　MIF 模块结构

Fig. 3　Structure of MIF module

MIF 模块的具体流程如算法 3 所示,其中输入特征 F 分别对应骨干网络中 C3、C4 和 C5 阶段的输出特征图,F_{Avg}、F_{Max} 分别表示平均池化和最大池化后的输出特征,$F_{AC[1]}$、$F_{AC[2]}$、$F_{AC[3]}$ 分别表示经过系数 r 为 1、2、3 的空洞卷积后的输出特征图,F_{Conv1} 表示经过 1×1 卷积后的输出特征图,BN 表示将通道特征进行归一化处理,$F_{Sigmoid1}$、$F_{Sigmoid2}$ 表示经过 Sigmoid 激活后得到的通道权重矩阵,F_{Cat} 表示将经过卷积操作得到的不同感受野的特征图进行拼接,F_{MIF} 表示该模块的最终输出特征。

算法 3　MIF 模块算法流程

输入:特征图序列 $F=(F_1,F_2,\cdots,F_n)$

输出:特征图序列 $F_{MIF}=(F_{MIF[1]},F_{MIF[2]},\cdots,F_{MIF[n]})$

1:for $i=1$ to n do

2:　$F_{Avg[i]}=\mathrm{Avg}(F_{[i]})$;　// $\mathrm{Avg}(F_{[i]})$ 表示进行平均池化

3:　for $j=1$ to n do

4:　$F_{AC[j]}=A\mathrm{Conv}_{[j]}(3\times3,F_{[i]})$;　// 表示进行核大小为 3×3 的空洞卷积

5:　end for

6:　$F_{Conv1[i]}=\mathrm{Conv}(1\times1,F_{[i]})$;

7:　$F_{Max[i]}=\mathrm{Max}(F_{[i]})$;　//Max ($F_{[i]}$) 表示进行最大池化

8:　$F_{Cat[i]}=\mathrm{Concat}(F_{AC[1]},F_{AC[2]},\cdots,F_{AC[n]},F_{Conv1[i]})$;　//在通道维度上进行拼接

9:　$F_{Sigmoid1[i]}=\mathrm{Sigmoid}(\mathrm{BN}(\mathrm{Conv}(1\times1,F_{Avg[i]})))$;　//表示经过 Sigmoid 函数激活

10:　$F_{Sigmoid2[i]}=\mathrm{Sigmoid}(\mathrm{BN}(\mathrm{Conv}(1\times1,F_{Max[i]})))$;

11:　$F_{MIF[i]}=F_{Cat}\times F_{Sigmoid1}+F_{Cat}\times F_{Sigmoid2}$;

12:end for

13:return F_{MIF}

1.2　可变形卷积特征金字塔网络(DCFPN)结构

传统的 FPN 在自上向下进行特征融合的过程中使用的是常规的卷积操作,但是固定的卷积核在对密集且不规则物体进行特征提取时,容易丢失一些包含物体局部信息的像素点,从而产生分类、回归不准确的问题。可变形卷积在常规卷积的基础上增加了空间偏移量,可以更好地适应图像中的空间变形。因此,本研究提出 DCFPN 结构,通过将可变形卷积融

入特征金字塔网络,使得模型在应对密集、形状不统一、尺寸不一致的复杂场景时能够很好地提取到物体特征,其结构如图 4 所示。

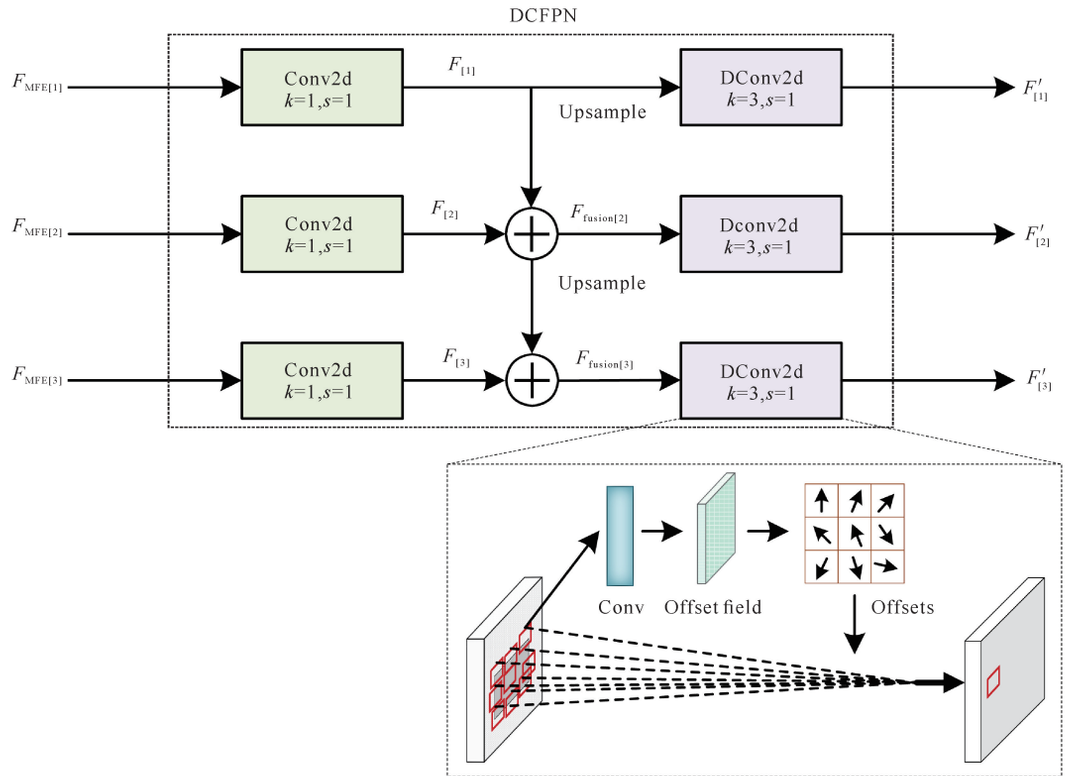


图 4　DCFPN 结构

Fig. 4　Structure of DCFPN

由图 4 可知,DCFPN 自上而下有 3 层。$F_{MFE[1]}$、$F_{MFE[2]}$、$F_{MFE[3]}$ 分别表示对应层中 MFE 模块的输出特征,$F_{[1]}$、$F_{[2]}$、$F_{[3]}$ 分别表示第一、第二、第三层 1×1 卷积的输出特征,$F_{fusion[2]}$、$F_{fusion[3]}$ 表示第二和第三层经过融合的特征,Upsample 表示对特征图进行上采样操作,本研究使用的是双线性插值上采样,DConv 表示可变形卷积,$F'_{[1]}$、$F'_{[2]}$、$F'_{[3]}$ 分别表示 DCFPN 结构第一、第二和第三层的输出特征,即分别对应网络中 P5、P4 和 P3 的输出。

DCFPN 结构的具体流程如算法 4 所示。

算法 4　DCFPN 结构算法流程

输入:特征图序列 $F_{MFE}=(F_{MFE[1]},F_{MFE[2]},F_{MFE[3]})$

输出:特征图序列 $F'=(F'_{[1]},F'_{[2]},F'_{[3]})$

1:for i=1 to 3 do

2:　　$F_{[i]}=\mathrm{Conv}(1\times1,F_{MFE[6-i]})$;

3:end for

4:$F_{fusion[2]}=F_{[2]}+\mathrm{Upsample}(F_{[1]})$;

5:$F_{fusion[3]}=F_{[3]}+\mathrm{Upsample}(F_{fusion[2]})$;

6:$F'_{[1]}=\mathrm{DConv}(3\times3,F_{[1]})$;

7:for i=2 to 3 do

8:　　$F'_{[i]}=\mathrm{DConv}(3\times3,F_{fusion[i]})$;

9:end for

10:return F'

2　实验

2.1　实验设置

本研究所有实验均在相同实验环境下进行,实验结果均为对应模型训练一次取最终结果。实验环境为 Ubuntu 18.04.6 LTS,处理器为 Intel Core i9-10980XE CPU @ 3.00 GHz×36,显卡为 GeForce RTX 4080,16 GB 显存。实验基于 PyTorch 深度学习框架,开发环境为 PyTorch1.13.0,Cuda 11.7,Python 版本为 3.7。训练时批大小设为 2,学习率设为 0.005,使用 SGD 优化器,输入 IMSE 与 FCOS 的图片大小为 600×600,输入其他模型的图片大小为 800×1 333。在 MS COCO 数据集[23]的对比实验中,IMSE 与 FCOS 的实验环境和实验结果的获取方式与上述一致。关于其他模型在 MS COCO 数据集上的实验结果,本研究引用了文献[24]中的相关数

据。为了保证公平性,本研究所选的对比模型均为目标检测领域中具有代表性的优秀模型,包括基于锚框和无锚框、单阶段和两阶段的检测算法,并考虑了各模型在不同数据集上的表现,以确保能够全面且深入地评估 IMSE 的优越性能。

2.2 数据集

为了验证本研究提出的 IMSE 在密集型的小目标检测场景下的适用性,本研究通过 3 个目标检测数据集进行验证,分别为 MS COCO[23]、CARPK[2] 和 WOOD 数据集,3 个数据集的图像数据信息如表 1 所示。

表 1 3 个数据集的图像数据信息

Table 1 Image data information of three datasets

数据集 Dataset	数据集划分 Partition of dataset	图片数量 Number of images	物体个数 Number of objects
MS COCO	Train	118 287	978 288
	Val	5 000	41 781
	Test	123 287	1 020 069
	Total	246 574	2 040 138
CARPK	Train	989	42 274
	Test	459	47 500
	Total	1 448	89 774
WOOD	Train	684	198 060
	Val	86	24 980
	Test	85	27 547
	Total	855	250 587

MS COCO 数据集[23] 包括训练集(Train,118 287 张图片)、验证集(Val,5 000 张图片)、测试集(Test,40 670 张图片),共有 80 个类别。数据集的数据特点是多目标、部分目标存在遮挡与噪声、包含各种尺寸的物体。CARPK[2] 数据集是一个通过无人机采集的大型公共数据集,该数据集包含从不同停车场捕获的近 90 000 辆汽车,是首个也是最大的支持物体计数的无人机视图数据集。数据集分为 Train(989 张图片)和 Test(459 张图片),共 1 个类别。数据集的特点是多目标、目标密集、目标尺寸小。WOOD 数据集由实际生产场景采集而得,该数据集由人工采集的 200 张图片,经过随机裁剪、水平翻转、镜像、图像锐化等数据增强操作得到了 855 张图片,其中 Train 有 684 张图片(包含有 198 060 根木头),Val 有 86 张图片(包含有 24 980 根木头),Test 有 85 张图片(包含有 27 547 根木头)。数据集的特点是图片中目标数量多、分布密集、尺寸多样、噪声多。

2.3 模型评价指标

本研究以 COCO 指标对模型进行性能评估,常见的评价指标有平均精度(Average Precision,AP)、AP_{75}、AP_S、AP_M、AP_L。AP 是通过计算 PR 曲线与横轴、纵轴之间的面积而得。PR 曲线是由精确率与召回率构成的曲线,其中横轴表示召回率,纵轴表示精确率。AP_{75} 表示 IoU 值(IoU 是预测的检测框和真实的检测框的交并比)为 0.75 时的平均精度,AP_S、AP_M、AP_L 分别表示模型检测小物体、中等物体、大物体的平均精度。平均精度越高,说明模型目标检测的准确性越高。

2.4 实验结果与分析

2.4.1 在 MS COCO 数据集上的对比实验

为了验证 IMSE 的有效性,表 2 为 IMSE 在 MS COCO 数据集上与主流目标检测模型的对比结果。其中,骨干网络 R50 与 R101 分别表示 ResNet50 与 ResNet101,Epochs 表示不同模型的迭代次数,AP、AP_{75}、AP_S、AP_M、AP_L 如上文所述均表示模型的平均精度,下表同。由表 2 可知,在该数据集的实验中,IMSE 在较少的 Epochs 下就能获得较高的 AP 值,很好地说明了网络结构设计的有效性。IMSE 的所有性能指标都优于基线 FCOS,AP 值从 47.6% 提高到 49.4%,AP_{75} 值从 52.1% 提高到 53.9%,证明了本研究提出的方法在多尺度目标检测方面的优越性。此外,IMSE 除了在 AP 和 AP_{75} 指标的表现上稍逊色于 CO-DETR,在 AP_S 指标上的性能稍逊色于基于 Transformer 的部分 DETR 系列模型,其余指标均表现最佳。本研究针对落后原因分析认为,这可能是由于 DETR 系列的模型在 Query 上提出一种新的建模方式,该方式借鉴了先验框的思想,以带有坐标信息的 Query 来辅助注意力机制的关注位置,因此在性能指标上比本研究提出的模型高,但该类模型也因先验框的引入导致在进行多尺度检测时,无法同时较好地识别不同尺度的物体,且可能带来计算复杂度高的问题。综上可知,IMSE 在非密集场景下的目标检测实现了综合性能的提升,验证了模型的泛化性和鲁棒性。

表 2　IMSE 与不同模型在 MS COCO 上的对比

Table 2　Performance comparison on MS COCO dataset

模型 Model	骨干网络 Backbone	迭代次数 Epochs	AP/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
DETR-DC5[25]	R101	500	44.9	47.7	23.7	49.5	62.3
Anchor-DETR-DC5[26]	R101	50	45.1	48.8	25.8	49.4	61.6
Efficient-DETR-DC5[27]	R101	36	45.7	49.5	28.2	49.1	60.2
Deformable-DETR[28]	R50	50	46.2	50.0	28.8	49.2	41.7
Dab-Deformable-DETR[29]	R50	50	46.9	50.8	30.1	50.4	62.5
YOLOv8	Darknet53	300	37.2	39.9	18.5	40.9	53.1
DN-Deformable-DETR[30]	R50	12	43.4	47.2	24.8	46.8	59.4
FCOS[10]	R101	12	47.6	52.1	24.7	54.1	71.1
CO-DETR[31]	R50	12	49.5	54.3	32.4	52.7	63.7
IMSE(ours)	R101	12	49.4	53.9	27.0	56.7	71.3

2.4.2　在 CARPK 数据集上的对比实验

如表 3 与图 5 所示，展示了 IMSE 与其他目标检测器在 CARPK 数据集上的实验结果。IMSE 在物体尺寸统一、物体排列密集的数据集上的所有性能指标均明显优于 YOLOv8、CO-DETR 等主流目标检测方法。具体而言，IMSE 在 AP 指标上达到了 75.8%，比基线提高了 1.4%，在 AP_S 指标上比基线提高了 1.1%。与 CO-DETR 模型相比，IMSE 在 AP 值上高出 21.7%，在更严格的 AP_{75} 指标上高出 16.1%，在 AP_S 指标上高出 16.1%。由此可见，IMSE 获得了更稳健的性能优势。图 5 直观地展现了 IMSE 在 CARPK 数据集上的准确率和召回率在所有对比模型中获得了最好的结果。可见，IMSE 不仅在多尺度检测方面占优，而且在处理物体大小和形状规则且密集的场景方面也表现出了良好的性能。

表 3　IMSE 与不同模型在 CARPK 上的对比

Table 3　Performance comparison on CARPK dataset

模型 Model	骨干网络 Backbone	迭代次数 Epochs	AP/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
DINO[32]	R50	100	48.1	55.1	7.0	49.4	—
FCOS[10]	R101	100	74.4	82.2	24.0	76.2	—
YOLOv7[33]	Darknet 53	100	60.9	75.1	5.8	62.8	—
YOLOv8	Darknet 53	100	57.4	72.1	2.7	59.2	—
CO-DETR[31]	R50	100	54.1	66.1	9.0	55.7	—
IMSE (ours)	R101	100	75.8	82.2	25.1	77.5	—

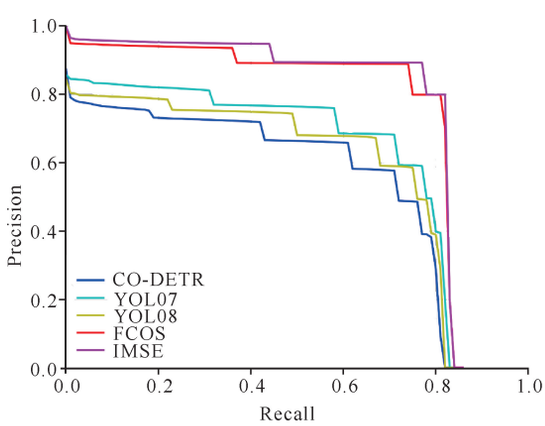


图 5　IMSE 与其他模型 PR 曲线对比

Fig. 5　PR curve comparison between IMSE and other models

2.4.3　在 WOOD 数据集上的对比实验

MS COCO 等数据集的数据特点均不属于图片中物体数目多、分布密集的特点，而在实际生产场景常常是物体数量多、尺寸变化大、分布密集，因此在 WOOD 数据集上验证 IMSE 的检测能力。表 4 是 IMSE 与不同模型在 WOOD 数据集上的性能对比。由表 4 可知，IMSE 在 AP、AP_{75} 与 AP_L 上均获得了所有对比模型中的最优结果，而 AP_S 未高于 DINO 与 YOLOv7。本研究分析认为，DINO 是基于传统多头自注意力机制进行训练的，相较于本研究的 WA 模块能更好地关注全局特征，而中小目标因存在遮挡等因素，需要建立全局联系才能更好地识别，因此本

研究在 AP_S 指标上要略逊色于 DINO。但是,由于 DINO 中使用了传统的多头自注意力(MHSA)模块,根据表 5 可知,在时间复杂度和硬件内存占用上,本研究的 WA 模块在参数量和计算量上更占优势。而 YOLOv7 中 SPPCSPC 结构通过 4 条不同的 MaxPool 分支增大感受野,且分支中池化核的尺寸较大,能够在更大范围内感知图像中的上下文信息,有助于理解小目标在整个图像中的位置和关系,因此在 AP_S 指标上略占优势。但是观察发现 YOLOv7 在其余性能指标上均表现不足,可能是由于最大池化操作只保留每个池化窗口中的最大值,导致部分重要信息丢失。综上,IMSE 在密集型的小目标检测场景下获得了更具有竞争性的性能表现。

表 4　IMSE 与不同模型在 WOOD 上的对比

Table 4　Performance comparison on WOOD dataset

模型 Model	骨干网络 Backbone	迭代次数 Epochs	AP/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
Mask RCNN[34]	R101	300	46.0	52.6	0.9	5.2	65.9
YOLOv5[35]	CSPDarknet	300	34.2	38.4	9.2	22.6	36.7
YOLOv7[33]	Darknet53	300	32.8	38.7	16.0	17.7	44.4
DINO[32]	R50	300	40.0	39.6	26.2	30.4	58.3
FCOS[10]	R101	300	52.9	62.3	13.3	43.8	65.6
IMSE(ours)	R101	300	55.0	62.8	14.8	38.5	67.7

表 5　WA 与 MHSA 的对比

Table 5　Comparison between WA and MHSA

模块 Module	每秒计算量/G FLOPs/G	参数量/M Parameters/M
WA	2.83	3.15
MHSA	3.90	14.75

为了更直观地验证 IMSE 在密集场景下小目标检测的优越性能,本研究在 WOOD 数据集上展示了不同模型的检测结果,如图 6 所示。通过观察发现,在真实的密集场景下进行检测时,Mask R-CNN 在某些目标存在遮挡情况时出现了遗漏和错误的检测,对于密集且较小的物体检测效果不佳;FCOS 虽然在密集场景下能较好地识别物体,但是在同一个物体中存在多个目标框重复出现和定位不够准确的问题;YOLOv5 与 YOLOv7 均存在目标框重复出现和无法较好地识别和定位到小目标的问题,特别是在目标之间存在明显重叠的情况下。这些问题可能是因为在面对密集的小目标物体时,仅依靠传统的卷积操作无法全面地学习和提取图像特征,而 IMSE 能够在复杂场景下具有较好地定位目标和噪声抑制的能力。

2.4.4　MHSA 模块与 WA 模块的计算量及参数量对比

MHSA 模块与 WA 模块的每秒计算量与参数量对比如表 5 所示。实验中两个模块均将注意力头数设置为 4,输入维度均为(1 024,30,30)。从实验结果可知,WA 模块在每秒计算量与参数量上都要少于 MHSA 模块。

2.4.5　消融实验

IMSE 主要包含 3 个功能模块:WA 模块、MIF 模块、DCFPN 结构。本研究的消融实验在 MS COCO、CARPK 和 WOOD 数据集上评估这 3 个模块的重要性。表中,√ 表示保留此组件,× 表示去除此组件。

表 6 是 IMSE 在 MS COCO 数据集上的消融实验结果,由实验结果可知在 WA 模块、MIF 模块和 DCFPN 结构都存在网络中时,模型 AP 达到了 49.4%;在仅添加 WA 模块与 DCFPN 结构时,模型 AP 为 48.5%;在仅添加 MIF 模块与 DCFPN 结构时,模型 AP 为 49.2%,且在 AP_L 指标上获得了最好的结果,本研究分析认为 MIF 模块通过不同大小的卷积和全局池化操作可以捕获到更大范围的上下文信息,而在加入了 WA 模块后在 AP_L 指标上性能略微有所下降,这可能是因为在引入 WA 模块后,因需要进行窗口注意力机制计算,模块之间的权重参数相互影响,网络更倾向于在更全面的性能提升上进行优化。由此验证了本研究所提出的 WA 模块、MIF 模块与 DCFPN 结构协同作用的有效性。

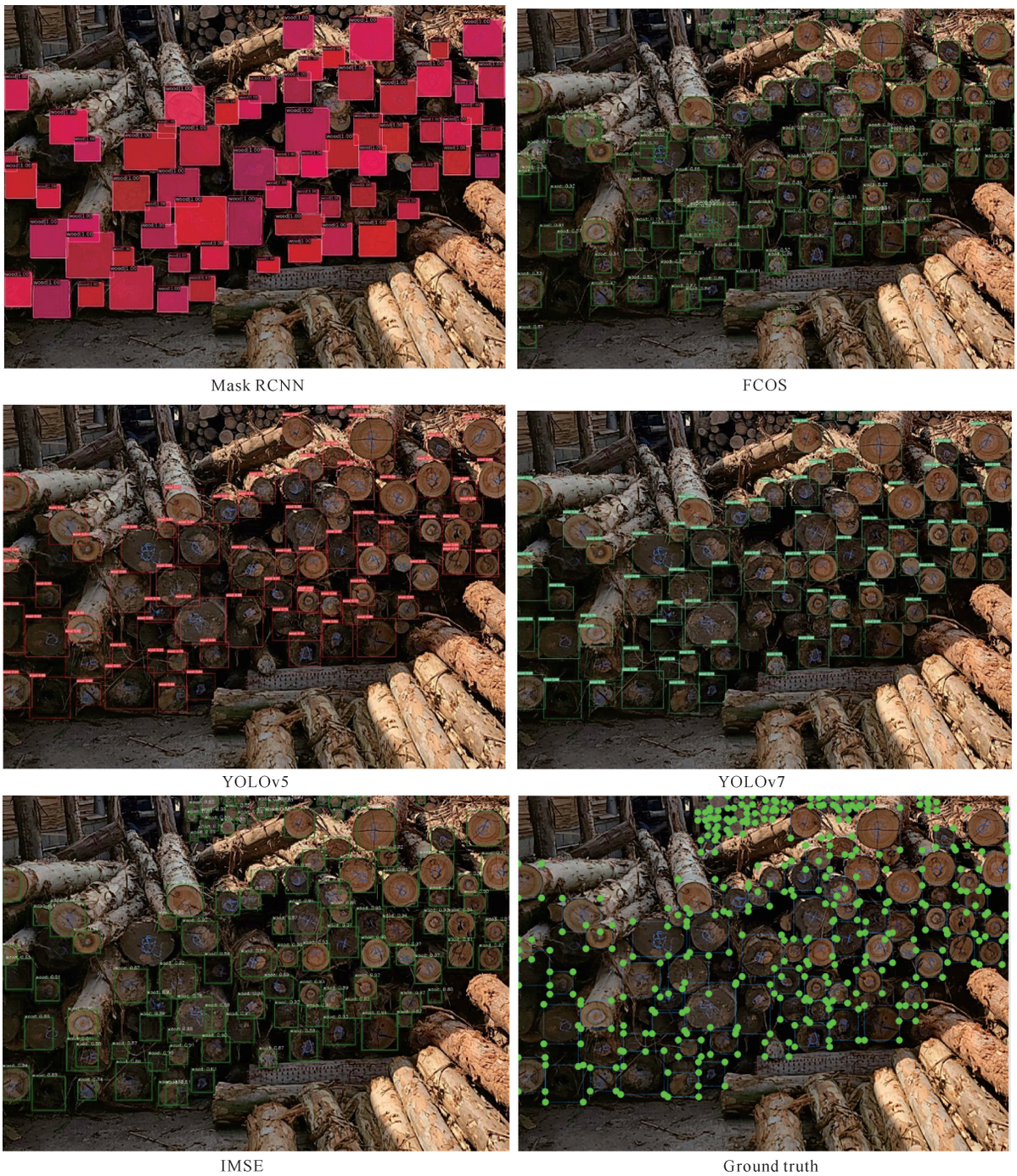


图 6　IMSE 与其他模型在 WOOD 数据集上检测效果展示

Fig. 6　Visualization of the detection effects of IMSE and other models on WOOD dataset

表 6　IMSE 在 MS COCO 数据集上的消融实验

Table 6　Ablation study of IMSE on MS COCO dataset

模型 Model	WA	MIF	DCFPN	AP/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
IMSE-0	×	×	×	38.1	40.7	16.3	41.7	60.7
IMSE-1	√	×	√	48.5	53.0	25.5	55.4	71.8
IMSE-2	×	√	√	49.2	53.8	25.7	56.4	72.5
IMSE-3	√	√	×	48.6	53.2	25.8	55.7	71.4
IMSE	√	√	√	49.4	53.9	27.0	56.7	71.3

表 7 是 IMSE 在 CARPK 数据集上的消融实验结果,由实验结果可知在 WA 模块、MIF 模块和 DCFPN 结构都存在网络中时,模型 AP 达到了 75.8%;在仅添加 WA 模块与 DCFPN 结构时,模型 AP 为 75.7%;在仅添加 MIF 模块与 DCFPN 结构时,模型 AP 为 75.1%。由此可知,模型引入 WA 模块、MIF 模块和 DCFPN 结构均对提高模型性能有益。

表 7 IMSE 在 CARPK 数据集上的消融实验

Table 7 Ablation study of IMSE on CARPK dataset

模型 Model	WA	MIF	DCFPN	AP/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
IMSE-0	×	×	×	74.4	82.2	24.0	76.2	—
IMSE-1	√	×	√	75.7	82.2	24.9	77.3	—
IMSE-2	×	√	√	75.1	82.2	23.5	76.8	—
IMSE-3	√	√	×	75.4	82.2	24.2	77.2	—
IMSE	√	√	√	75.8	82.2	25.1	77.5	—

表 8 是 IMSE 在 WOOD 数据集上的消融实验结果。由实验结果可知在 WA 模块、MIF 模块和 DCFPN 结构都存在网络中时,模型 AP 达到了 55.0%;在仅添加 WA 模块与 DCFPN 结构时,模型 AP 为 55.2%;在仅添加 MIF 模块 DCFPN 结构时,模型 AP 为 54.8%。由实验结果可知,WA 模块、MIF 模块与 DCFPN 结构均同时存在网络中时,只有 AP_S 指标达到了最优值。针对这个情况,本研究认为这是由于 WA 模块的作用是关注特征图的全局空间信息,MIF 模块的作用是关注特征图的局部空间信息和通道信息,两者均对空间信息进行关注。当 WA 与 MIF 模块同时作用于网络时,若不区分两个模块输出特征的重要程度而直接进行特征融合,会导致网络无法很好地表达关键区域的特征,从而小幅度影响模型的性能。但总体来说,模型引入 WA 模块、MIF 模块和 DCFPN 结构均对提高模型的性能有益。

表 8 IMSE 在 WOOD 数据集上的消融实验

Table 8 Ablation study of IMSE on WOOD dataset

模型 Model	WA	MIF	DCFPN	AP/%	AP_{75}/%	AP_S/%	AP_M/%	AP_L/%
IMSE-0	×	×	×	52.9	62.3	13.3	43.8	65.6
IMSE-1	√	×	√	55.2	62.9	13.9	46.8	67.8
IMSE-2	×	√	√	54.8	62.9	14.2	46.6	67.3
IMSE-3	√	√	×	53.6	61.4	14.5	26.9	67.1
IMSE	√	√	√	55.0	62.8	14.8	38.5	67.7

3 结论

本研究主要探讨了现有目标检测模型在多尺度特征提取中存在的不足,设计了 MFE 模块和 DCFPN 结构。在 MFE 模块中,WA 模块通过不同大小的滑动窗口对特征图进行自注意力机制计算,在基本不降低模型性能的同时大大降低了模型的计算量;MIF 模块能够同时兼顾特征图的局部空间和通道特征信息,有效提高了模型的特征提取能力。DCFPN 结构在应对密集、形状不一和尺寸不一致的物体检测场景时能够很好地提取到物体特征,以解决网络在处理复杂检测场景时出现的漏检、错检和重检的问题。最终,在 MFE 模块与 DCFPN 结构的基础上提出了密集型小目标检测网络 IMSE。通过在 MS COCO、CARPK、WOOD 3 个不同类型的数据集上进行对比实验和消融实验,证明了本研究提出的 IMSE 在密集场景下的小目标检测中有着较优异的性能表现。

参考文献

[1] HAN T,BAI L,GAO J Y,et al. DR. VIC:decomposition and reasoning for video individual counting [C]//Proceedings of the IEEE/CVF Conference on Computer Vi-

sion and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 3083-3092.

[2] HSIEH M R, LIN Y L, HSU W H. Drone-based object counting by spatially regularized regional proposal network [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 4145-4153.

[3] WANG L Q, YANG J Y, ZHANG Y F, et al. Depth-aware concealed crop detection in dense agricultural scenes [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 17201-17211.

[4] 杨攀,郑积仕,冯芝清,等. 基于 Mask R-CNN 的密集木材检测分割方法[J]. 林业工程学报, 2022, 7(2): 135-142.

[5] 霍爱清,张书涵,杨玉艳,等. 密集交通场景中改进 YOLOv3 目标检测优化算法[J]. 计算机工程与科学, 2023, 45(5): 878-884.

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

[7] WU Z W, WANG X F, JIA M, et al. Dense object detection methods in RAW UAV imagery based on YOLOv8 [J]. Scientific Reports, 2024, 14(1): 18019.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2017: 6000-6010.

[9] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with transformers [C]//European Conference on Computer Vision. Cham: Springer, 2020: 213-229.

[10] TIAN Z, SHEN C H, CHEN H, et al. FCOS: a simple and strong anchor-free object detector [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(4): 1922-1933.

[11] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 2117-2125.

[12] NAYAK A, BOUAZIZI M, AHMAD T, et al. Evaluation of fully convolutional one-stage object detection for drone detection [C]//International Conference on Image Analysis and Processing. Cham: Springer, 2022: 434-445.

[13] FU H Y, YUE Y K, WANG W, et al. Ramie plant counting based on UAV remote sensing technology and deep learning [J]. Journal of Natural Fibers, 2023, 20(1): 2159610.

[14] LIU S, CHI J N, WU C D. FCOS-lite: an efficient anchor-free network for real-time object detection [C]//2021 33rd Chinese Control and Decision Conference (CCDC). Piscataway, NJ: IEEE, 2021: 1519-1524.

[15] YU J B, CHENG X, LI Q F. Surface defect detection of steel strips based on anchor-free network with channel attention and bidirectional feature fusion [J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-10.

[16] XIE J X, ZHANG X W, LIU Z Q, et al. Detection of Litchi leaf diseases and insect pests based on improved FCOS [J]. Agronomy, 2023, 13(5): 1314.

[17] GAO P, TIAN T, ZHAO T M, et al. Double FCOS: a two-stage model utilizing FCOS for vehicle detection in various remote sensing scenes [J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 4730-4743.

[18] PAVEZ L, SAAVEDRA J M. NL-FCOS: improving FCOS through non-local modules for object detection [C]//2022 26th International Conference on Pattern Recognition (ICPR). Piscataway, NJ: IEEE, 2022: 4651-4657.

[19] WANG Y N, LIN X N, ZHANG X S, et al. Improved FCOS for detecting breast cancers [J]. Current Medical Imaging, 2022, 18(12): 1291-1301.

[20] ZHOU B, LI B, LAN W F, et al. SDH-FCOS: an efficient neural network for defect detection in urban underground pipelines [J]. Computers, Materials & Continua, 2024, 78(1): 633-652.

[21] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 764-773.

[22] YU F, KOLTUN V, FUNKHOUSER T. Dilated residual networks [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2017: 472-480.

[23] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//European Conference on Computer Vision. Cham: Springer, 2014: 740-755.

[24] ZHAO Y, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection [C]//2024 IEEE/

CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2024: 16965-16974.

[25] CHEN Q, CHEN X K, WANG J, et al. Group DETR: fast DETR training with group-wise one-to-many assignment [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2023: 6633-6642.

[26] WANG Y M, ZHANG X Y, YANG T, et al. Anchor DETR: query design for transformer-based detector [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2022, 36(3): 2567-2575.

[27] YAO Z Y, AI J B, LI B X, et al. Efficient DETR: improving end-to-end object detector with dense prior [EB/OL]. (2021-04-03)[2024-05-21]. https://doi.org/10.48550/arXiv.2104.01318.

[28] ZHU X Z, SU W J, LU L W, et al. Deformable DETR: deformable transformers for end-to-end object detection [EB/OL]. (2021-03-18)[2024-05-21]. https://doi.org/10.48550/arXiv.2010.04159.

[29] LIU S L, LI F, ZHANG H, et al. Dab-DETR: dynamic anchor boxes are better queries for DETR [EB/OL]. (2022-03-30)[2024-05-21]. https://doi.org/10.48550/arXiv.2201.12329.

[30] LI F, ZHANG H, LIU S L, et al. DN-DETR: accelerate DETR training by introducing query denoising [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2022: 13619-13627.

[31] ZONG Z, SONG G, LIU Y. DETRs with collaborative hybrid assignments training [C]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2023: 6748-6758.

[32] ZHANG H, LI F, LIU S L, et al. DINO: DETR with improved denoising anchor boxes for end-to-end object detection [EB/OL]. (2022-07-11)[2024-05-21]. https://doi.org/10.48550/arXiv.2203.03605.

[33] WANG C Y, BOCHKOVSKIY A, LIAO H Y M. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE, 2023: 7464-7475.

[34] HE K M, GKIOXARI G, DOLLÁR P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE, 2017: 2961-2969.

[35] YANG G H, FENG W, JIN J T, et al. Face mask recognition system with YOLOV5 based on image recognition [C]//2020 IEEE 6th International Conference on Computer and Communications (ICCC). Piscataway, NJ: IEEE, 2020: 1398-1404.

An Intensive Small Object Detection Network Based on Multi-scale Feature Extraction

YUAN Chang'an[1,2], WANG Wenji[1], HUANG Haojie[3], QIN Zhengyou[1], ZHANG Jinyong[1], LIAO Huixian[4], QIN Xiao[1,5**], LI Xiaosen[6], LI Yongyu[1], FU Yunqin[1], TAN Sijing[1], QIAN Quanmei[1], WU Kunsheng[7]

(1. Guangxi Key Laboratory of Human-Computer Interaction and Intelligent Decision Making, Nanning Normal University, Nanning, Guangxi, 530100, China; 2. Guangxi Academy of Sciences, Nanning, Guangxi, 530007, China; 3. Guangxi Technical Service Company, China Communications Services Corporation Limited, Nanning, Guangxi, 530000, China; 4. College of Digital Technology, Guangdong Vocational College of Finance and Trade, Qingyuan, Guangdong, 511510, China; 5. Guangxi Regional Collaborative Innovation Center for Multi-Source Data Integration and Intelligent Processing, Guilin, Guangxi, 541004, China; 6. School of Artificial Intelligence, Guangxi Minzu University, Nanning, Guangxi, 530006, China; 7. Nanning Arboretum, Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530225, China)

Abstract:Aiming at the problem that existing object detection algorithms without anchor frame are difficult to extract multi-scale target features effectively in dense scenes,an Intensive small object detection network based on Multi-Scale feature Extraction (IMSE) is proposed. Firstly,a multi-scale feature enhancement module including a Window Attention (WA) module and a Multi-scale Information Fusion (MIF) module is proposed. The global context connection is established to enhance the feature expression of IMSE in dense scenes,which enables more effective extraction of multi-scale features of detection objects. Secondly,a Deformable Convolutional Feature Pyramid Network (DCFPN) structure is proposed,which introduces dilated convolution for feature enhancement,thereby effectively improving the ability of IMSE to detect irregularly shaped and irregularly distributed objects. Finally,the fused multi-scale features are input into the detection head for classification and bounding box regression tasks. IMSE was then validated on the public datasets MS COCO and CARPK and the WOOD dataset constructed based on actual production scenarios. The experimental results showed that the Average Precision (AP) of IMSE on the three datasets reached 49.4%,75.8%, and 55.0%,respectively,which were 1.8%,1.4%,and 2.1% higher than that of the original FCOS method, verifying the effectiveness of the proposed model.

Key words:object detection;self-attention mechanism;feature pyramid network;dilated convolution;deformable convolution

责任编辑:陆　雁