基于注意力机制的 YOLOv5 优化模型

潘烨新[1,2],黄启鹏[1,2],韦 超[1,2],杨 哲[1,2]
(1. 苏州大学 计算机科学与技术学院,江苏 苏州 215006;
2. 省计算机信息处理技术重点实验室,江苏 苏州 215006)

摘 要:目标检测是机器视觉研究中的重要分支。目前在工业生态中应用广泛的 YOLOv5 模型经过版本迭代,在预测权重大小以及检测精度方面都有所优化,但模型的处理速度仍然较低,尤其是对于小目标及遮挡目标的检测效果有待改进。该文提出一种基于注意力机制的 YOLO v5 改进模型。首先,通过引入维度关联注意力机制模块进行特征融合,提升主干网络的特征提取能力,达到改善小目标与遮挡目标的检测效果;其次,采用 SIoU 损失函数代替 CIoU 损失函数,作为新的边界框回归参数的损失函数,提高边界框的定位精度以及检测速度。实验结果显示,优化模型的平均精度均值达到 87.8%,相比于 YOLOv5 提高了 4.7 百分点,在单 GPU 上模型的检测速度达到 83.3 FPS。

关键词:机器视觉;深度学习;目标检测;注意力机制;损失函数
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)12-0163-08
doi:10.3969/j.issn.1673-629X.2023.12.023

YOLOv5 Optimization Model Based on Attention Mechanism

PAN Ye-xin[1,2],HUANG Qi-peng[1,2],WEI Chao[1,2],YANG Zhe[1,2]
(1. Department of Computer Science and Technology,Soochow University,Suzhou 215006,China;
2. Jiangsu Provincial Key Laboratory for Computer Information Processing Technology,Suzhou 215006,China)

Abstract:With the development of machine vision technology,target detection has become an important branch. At present,the YOLOv5 model,which is widely used in the industrial ecology,has undergone version iterations and has been optimized in terms of prediction weight and detection accuracy,but the processing speed of the model is still not high,especially for small targets and occluded objects. The detection effect needs to be improved. We propose an improved model of YOLO v5 based on attention mechanism. First of all,by introducing the dimension related attention mechanism module for feature fusion,the feature extraction ability of the backbone network is improved to improve the detection effect of small targets and occluded objects; secondly,the SIoU loss function is used instead of the CIoU loss function as a new bounding box regression parameter. The loss function improves the positioning accuracy and detection speed of the bounding box. The experimental results show that the average precision of the optimized model reaches 87.8%,which is 4.7 percentage points higher than that of YOLO v5,and the detection speed of the model on a single GPU reaches 83.3 FPS.

Key words:computer vision;deep learning;object detection;attention mechanism;loss function

0 引 言

目标检测是机器视觉领域重要的研究内容之一[1],目前主流的检测模型分为单阶段模型、双阶段模型以及基于 Transformer 解编码结构的模型[2]。双阶段算法先提取候选区域再进行分类和回归,如 RCNN[3],Faster R-CNN[4] 系列。这些方法在检测精度上表现出色,但由于计算量较大,检测速度较慢。单阶段检测算法无需提取候选区域,2直接对每个特征图进行回归预测。经典的单阶段检测算法有 YOLO[5],SSD[6],FCOS[7] 等系列算法,YOLO 因检测速度快被广泛应用于工业和日常生活中。但由于 YOLOv5 使用的骨干网络 CSPDarknet-53[8] 提取的特征图尺寸较小、分辨率较低、像素感受野较大,导致小目标的定位性能较差,因此整体性能仍存在一定的优化空间。同时整个网络中主要负责提取图像特征的是 C3 模块,分布在网络的骨干和颈部中。在骨干部分,C3 模块可

收稿日期:2023-03-09 **修回日期**:2023-07-12
基金项目:国家自然科学基金资助项目(62002253);教育部产学合作协同育人项目(220606363154256);国家级大学生创新创业训练计划项目(202210285042Z)
作者简介:潘烨新(1999-),男,硕士研究生,研究方向为计算机视觉与深度学习;通信作者:杨 哲(1978-),男,副教授,博士,研究方向为人工智能、大数据等。

以为特征图提取到大量的位置与细节信息,但语义信息提取的较少。当特征图前向传播到颈部部分后,在特征金字塔网络(FPN[9])与像素聚合网络(PAN)框架的结合作用下,C3 模块主要负责纹理特征的提取,此时会获得较为丰富的语义信息,但丢失了大量的位置与细节信息。导致网络模型对于小物体及有遮挡目标产生漏检误检情况,性能下降。该文提出一种基于注意力机制的 YOLOv5 优化模型。通过引入 DRA(Dimension Related Attention,维度关联注意力)模块来解决 C3 模块信息丢失问题,增强主干网络提取图像特征的能力;针对感受野大而导致的定位困难问题,引入新的定位计算损失函数,在提高边界框的定位精确度的同时优化模型的推理速度,间接提升模型的性能。在通用数据集上的实验结果表明,该方法提升了主干网络的特征提取能力,降低了回归参数的损失,从而提升了模型的整体性能。

1 相关工作

1.1 目标检测模型

双阶段模型的代表 RCNN 开创性地使用深度学习模型进行目标检测。但存在两个问题:一是经过缩放处理后会使一些图片特征信息丢失,从而降低检测的准确性,不利于小目标的检测;二是在训练和预测中,RCNN 的速度都非常慢。Faster R-CNN 提出了区域生成网络(Region Proposal Networks,RPN)用于提升检测框的生成速度,最终精度较高,但实时性与检测小目标的效果差。YOLO 是单阶段模型的起始作,不再生成候选区而是直接进行分类和回归。v1 通过将图像划分成多个网格来生成候选框。相比于二阶段模型,检测速度有了很大提高,但精度相对较低,尤其在小目标检测方面。v2[10]改变了主干网络,相比 v1 模型在精度、速度和分类数量上都有了很大的改进,但由于每个网格只能预测一个物体,当同一个网格内包含多个物体时只能检测到一个,因此对小物体的识别效果仍然非常差。v3[11]中提出了基于锚框的思想,使得最后的特征图上基于每个单元格都有三个不同的尺寸大小的锚框,进而对锚框进行分类与回归。v4[8]针对预处理以及激活函数问题,分别引入了 Mosaic 数据增强手段以及 Mish 激活函数[12],使得网络的收敛速度与精度进一步提升,但仍然存在框定位不准以及召回率低的问题。YOLOv5 在对模型主干以及颈部的基础改进之外,更换了新的损失函数计算方法,同时优化了一直存在的正负样本分配问题。但对于整体而言,预测框的回归精度与速度仍然较差。研究者们针对不同应用场景和问题,提出了基于 YOLOv5 的一系列应用优化算法。张浩等人[13]提出的算法旨在提高无人机视角下密集小目标的检测精度,并保证实时性。李永军等人[14]将红外成像与 v5 模型相结合,解决动态识别与密集目标的问题。窦其龙[15]通过优化深度学习网络、重新设置锚点框大小和嵌入 GDAL 模块,提高检测速度和降低漏检率。刘闪亮[16]则提出了注意力特征融合结构,进一步提高模型对小目标的检测性能。田枫[17]提出了 Cascade-YOLOv5,用于油田场景规范化着装检测,来提高检测性能。这些算法都是基于 YOLOv5 的改进和优化,以适应不同领域和应用需求。

1.2 注意力机制

在机器视觉领域,常使用的是软注意力,对其按维度可划分为通道注意力、空间注意力和自注意力。通道注意力旨在联系不同特征图,通过网络训练获取每个通道的重要度从而赋予不同权重最终强化重要特征,代表模型如 SE-Net(Squeeze and Excitation)[18]。空间注意力通过空间转换和掩码加权等方式增强兴趣区域[19]的同时弱化背景区域。如轻量级注意力模块 CBAM[20]。自注意力旨在最大化利用特征自身的固有信息进行交互。在 Google 提出的 Transformer 架构中被实际应用,何凯明等人将其应用到 CV 领域,并提出了 Non-Local 模块[21],通过自注意力机制有效地捕获长距离的特征依赖,实现全局上下文信息的建模。注意力机制模块众多,模型性能差异大,对比评估一些新型且有效的注意力机制模块,并进行一些创新改进,对提升复杂多尺度目标的检测性能是非常有意义的。

2 改进后的 YOLOv5 优化模型

2.1 DRA 注意力机制

DRA 模块在经典的 SE 模块上做出优化,如公式 1 所示,它可以对网络中任意的中间特征张量进行转化变换后输出同样尺寸的张量。DRA 模块结构如图 1 所示。

$$\boldsymbol{X}=[x_1,x_2,\cdots,x_c]\in R^{H\times W\times C}\rightarrow$$
$$\boldsymbol{Y}=[y_1,y_2,\cdots,y_c]\in R^{H\times W\times C} \tag{1}$$

在原先同时关注空间和通道信息的基础上,通过改变全局池化的操作,保留通道间信息的同时考虑重要的空间信息。

通道注意力常采用全局池化编码全局空间信息,简而言之是全局信息被压缩成了一个标量,而压缩完之后的标量难以保留重要的空间信息。为解决此问题,DRA 将全局池化操作改进为两个 1 维向量的编码操作。

为了获取输入图像的高度与宽度上的注意力,并完成对精确位置信息的编码,对于输入特征图,使用池化核 $(1,W)$ 和 $(H,1)$ 分别对高度和宽度的特征进行平均池化,从而获得两个方向的特征图,如式 2 和式 3

所示。

$$z_c^h(h)=\frac{1}{W}\sum_{0\leqslant i\leqslant W}x_c(h,i) \tag{2}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\leqslant j\leqslant H}x_c(j,w) \tag{3}$$

对比全局池化的压缩方式,这样能够允许注意力模块捕捉单方向上的长距离关系,同时保留另一个方向上的空间信息,帮助网络模型更准确地定位目标。

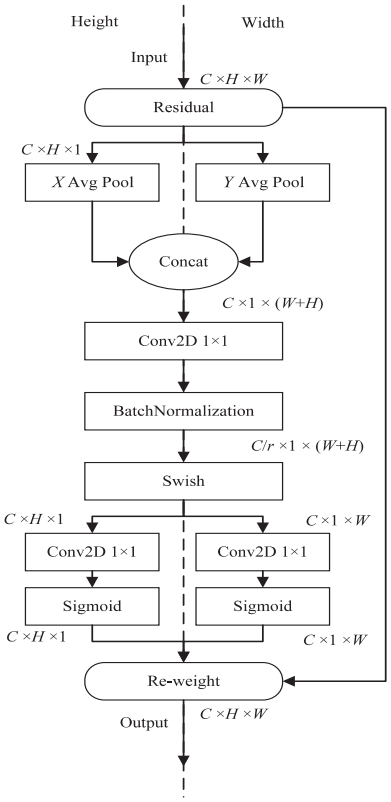


图 1　DRA 注意力机制

接着将获得全局感受野的高度和宽度两个方向的特征图按通道维度拼接在一起,主要目的是方便之后进行批量归一化(Batch Normalization,BN)操作。将它们送入卷积核为 1 × 1 的共享卷积模块 Conv2D,将其维度降低为 C/r ,r 为可设定的缩减因子,接着对其进行 BN 处理,将得到的特征图记为 F_1,最后送入 Swish 激活函数进行非线性变换,将这种变换记为 δ ,即可得到尺寸为 $C/r \times 1 \times (W+H)$ 的包含横向和纵向空间信息的特征图 f,如公式 4 所示。

$$f=\delta(F_1([Z^h,Z^w])) \tag{4}$$

随后将 f 按照原来的高度和宽度进行卷积核大小为 1×1 的卷积,分别得到通道数与原来一样的两个独立的特征 f^h 和 f^w ,最后经过 Sigmoid 激活函数后,分别得到特征图在高度上的注意力权重 g^h 和在宽度方向的注意力权重 g^w ,如式 5 和式 6 所示。

$$g^h=\sigma(F_h(f^h)) \tag{5}$$

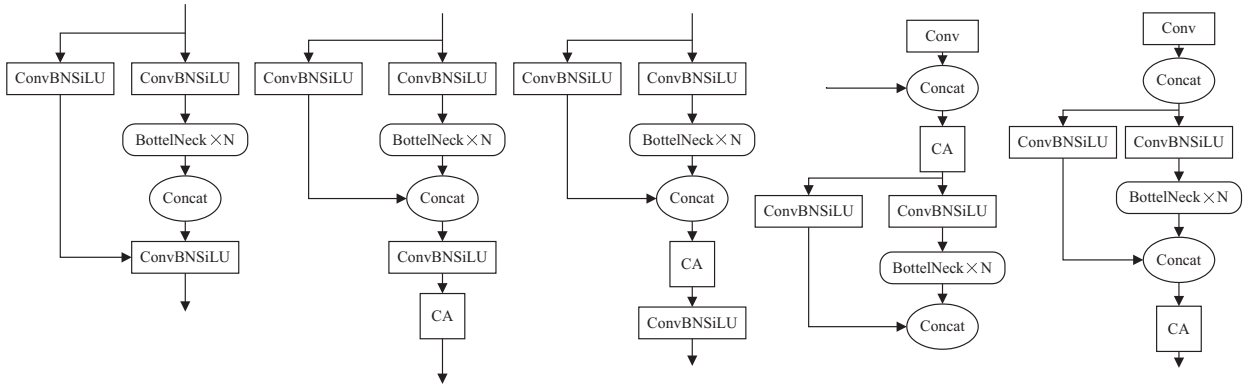
$$g^w=\sigma(F_w(f^w)) \tag{6}$$

最后在原始特征图上通过乘法加权计算,得到最终在宽度和高度方向上带有注意力权重的特征图,如式 7 所示。

$$y_c(i,j)=x_{c(i,j)}\times g_c^h(i)\times g_c^w(j) \tag{7}$$

2.2　注意力机制融合

针对原模型对于特征表达能力的不足,不易识别难检目标,以及由于只考虑通道信息而缺失方向相关信息带来的定位不准等问题,通过将 DRA 模块插入到网络模块中解决。同时由于原模型的主干,颈部,检测头三层结构会带来结果的干扰性以及不确定性,该文分析了在三层结构不同位置插入 DRA 模块的效果。如图 2 所示,分别在主干,颈部,预测头中插入 DRA 模块。对于主干部分,细分了 DRA 的插入位置。根据后续实验表 1 的数据,最终确定选择(b)方式插入到主干,将新的整体结构命名为 CDRA 模块,取代原模型主干中的 C3 模块。CDRA 模块相比 YOLO v5 原先的 C3 模块最大的改进在于,每个权重都包含了通道间信息、横向空间信息和纵向空间信息,能够帮助主干网络更准确地定位目标信息,增强识别能力。



(a)原 C3 结构　(b)插入主干 C3 最后一层　(c)插入主干 C3 的残差模块　(d)插入 Neck 部分　(e)插入 Head 部分

图 2　注意力插入位置对比

将运用维度关联注意力机制融合的 CDRA 替换原 C3 模块,经过改进后的 YOLOv5s 结构如图 3 所示。图 3 中,YOLOv5s 主要由主干网络、颈部、预测头部网络三部分组成,主干部分的替换工作对改进后的

YOLOv5s 性能提升起到决定性作用。

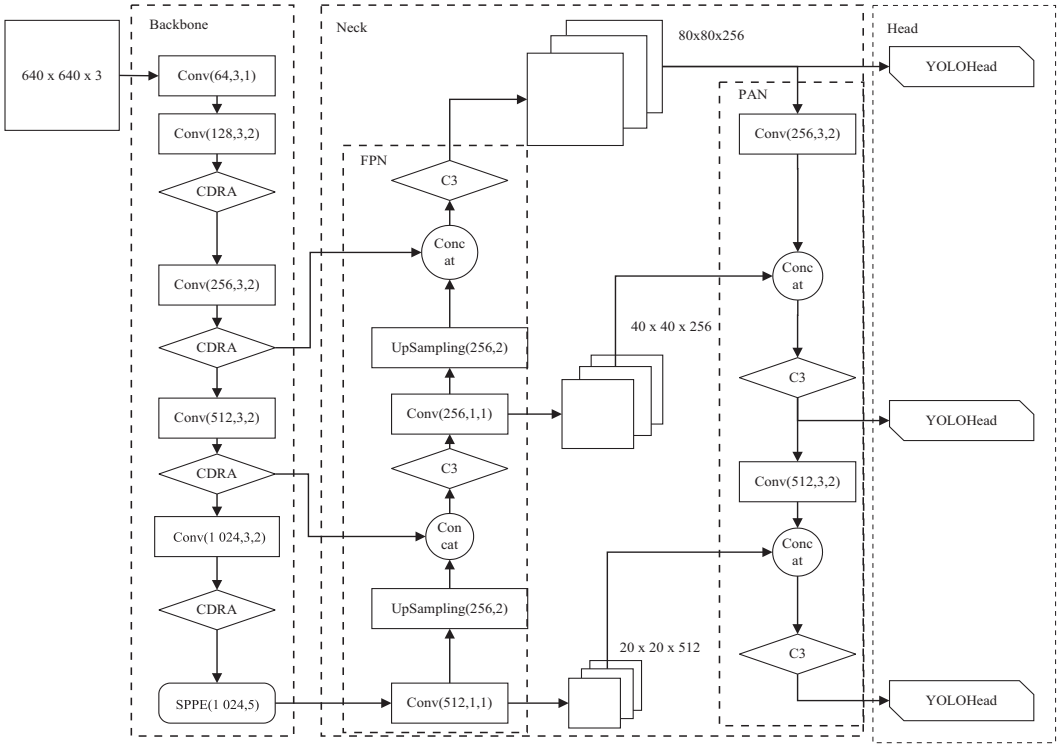


图 3 改进 YOLOv5s 网络

2.3 损失函数

目标检测模型的损失函数通常由三个部分构成，分别为预测框的定位损失 L_{box}，置信度损失 L_{obj}，分类损失 L_{cls}，整体的网络损失的计算如式 8 所示。

$$L = L_{box} + L_{obj} + L_{cls} \tag{8}$$

其中置信度损失和分类损失均采用交叉熵损失(Binary Cross Entropy Loss)，公式如式 9 所示。

$$\text{Loss} = \frac{1}{n}\sum_{i}^{n}\left[y_i \cdot \log(\sigma(x_i)) + (1 - y_i) \cdot \log(1 - \sigma(x_i))\right] \tag{9}$$

预测框的定位损失用来衡量当前模型所给出的预测框与真实框之间位置上的误差，具体会计算两者的中心坐标、高、宽等误差。早期模型一般采用 L1，L2，smooth L1 来计算该损失，但其忽略了 4 个回归参数之间的相关性。当前常用的是交并比损失(Intersection over Union，IoU)，IoU 的计算公式如式 10 所示。

$$\text{IoU} = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \tag{10}$$

其中，$B = (x,y,w,h)$ 表示预测框的位置，$B^{gt} = (x^{gt}, y^{gt}, w^{gt}, h^{gt})$ 表示真实框的位置。

IoU 损失的计算公式如式 11 所示。

$$\text{loss}_{\text{IoU}} = 1 - \text{IoU} \tag{11}$$

由公式可知两个矩形框重合度越高，IoU 越接近 1，则损失越接近 0。但采用 IoU 损失生效的情况仅在两框之间有重叠的部分，对于非重叠的两个框，IoU 损失不会提供任何可供传递的梯度。

YOLOv5 原始模型中采用 CIoU 作为边界框的定位损失函数。CIoU 是在 DIoU(Distance IoU)[22] 的基础上考虑了两框的长宽比而演化而来，但是仍然没有考虑到真实框与预测框之间不匹配的方向。这种不足导致 CIoU 收敛速度较慢且效率较低。

为了解决 CIoU 存在的问题，该文引入 SIoU[23] 用以改进，保留了原损失函数的全部性质，同时考虑方向框的角度回归问题，重新定义了惩罚指标。

SIoU 由四部分组成：角度损失 Λ、距离损失 Δ、形状损失 Ω 以及交并比损失(IoU)。

角度损失函数组件 Λ，如式 12 所示。

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{12}$$

其中，x 是直角三角形中的对边比斜边，如图 4 所示，α 是两框中心连线与预测框中心水平线的夹角。则 x 可由式 13 表示。

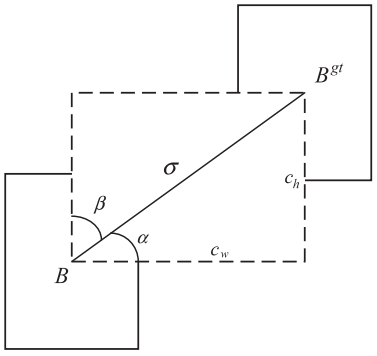


图 4 损失函数组件示意图

$$x = \frac{c_h}{\sigma} = \sin\alpha \tag{13}$$

其中,c_h 为真实框和预测框中心点的高度差,σ 为真实框和预测框中心点的距离,可由式 14 表示。

$$\sigma = \sqrt{(b_{c_x}^{gt} - b_{c_x})^2 + (b_{c_y}^{gt} - b_{c_y})^2} \tag{14}$$

将式 13 带入式 12 化简可得最终的角度损失计算公式,如式 15。

$$\Lambda = \sin(2\alpha) \tag{15}$$

可见角度损失本质就是计算两倍夹角的正弦值。所以当 α 为 0 或 $\frac{\pi}{2}$ 时,角度损失为 0,在训练过程中若 $\alpha < \frac{\pi}{4}$,则需要最小化 α,否则需要最小化 β。

定义角度损失后,考虑到当出现同时存在一个角度很小但是很近,与一个角度很大但是很远的框的情况时,近的框总是会被优先选择,所以直接使用角度损失不合理,还需要考虑距离与角度的互相关系。为了保证距离与角度的平衡,将角度损失同时考虑,重新定义了距离损失函数,记为 Δ,如式 16 所示。

$$\Delta = \sum_{t=x,y} (1 - e^{-\gamma\rho_t}) \tag{16}$$

其中,ρ_x,ρ_y,γ 定义如式 17 所示。

$$\rho_x = \left(\frac{b_{c_x}^{gt} - b_{c_x}}{c_w}\right)^2, \rho_y = \left(\frac{b_{c_y}^{gt} - b_{c_y}}{c_h}\right)^2, \gamma = 2 - \Lambda \tag{17}$$

ρ_t 是使用原始的距离损失的平方次幂来赋权重,说明距离的影响要大于角度的影响。

形状损失主要负责从长宽角度评价预测框的回归参数与真实框是否相似,记为 Ω,如式 18 所示。

$$\Omega = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta} \tag{18}$$

其中,ω_w,ω_h 如式 19 所示。

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{19}$$

θ 用来控制整体对形状损失的关注程度。

综合考虑上述 3 项以及默认的 IoU 损失,就可以得到最后的预测框的定位损失函数,如式 20 所示。

$$L_{\text{box}} = 1 - \text{IoU} + \frac{\Delta + \Omega}{2} \tag{20}$$

3　实验与分析

3.1　实验环境

实验环境配置:Window10 操作系统,32 核 Intel CPU,32 GB 内存,两块 TELSA A100,40 GB 存储空间。深度学习框架为 PyTorch1.10,图形处理器驱动为 CUDA11.4 和 Cudnn8。训练过程中所使用的优化器为 Adam[24],初始学习率为 0.01,动量因子为 0.937,权重衰减为 0.000 25,批尺寸为 32,总迭代次数设置为 300。

3.2　数据集及预处理

使用 Pascal VOC07+12 训练集以及 VOC07 测试数据集来评估模型性能,包含 20 个类别的常见交通工具、家具和动物等图像,可用于目标检测任务。共包含 8 281 张训练图像、8 333 张验证图像和 4 952 张测试图像。同时,在 ImageNet 数据集上对模型的主干网络进行了预训练,在训练过程中,使用 Mosaic 数据增强技术对前 75% 的训练周期进行了处理。

3.3　评估指标

使用检测速度、检测精度和损失函数收敛曲线等客观指标来评价模型的性能。其中,FPS 是检测速度的评价指标。AP(Average Precision)是指在 0 ~ 1 范围内 P(Precision,正确率)指标对 R(Recall,召回率)指标的积分,即 P-R 曲线下面积,AP 值越大,模型精度越高。mAP 是平均精度均值,指的是每个目标类别 AP 的均值。

计算公式分别如式 21 ~ 24 所示。

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{21}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{22}$$

$$\text{AP} = \int_0^1 P(R)\,\mathrm{d}R \tag{23}$$

$$\text{mAP} = \frac{1}{N}\sum_{i=1}^{N} \text{AP}_i \tag{24}$$

式中,TP 表示正确识别的目标数量,FP 表示识别错误的目标数量,FN 表示未被识别出目标数量。如果 IoU 大于一定阈值,则检测框被标记为 TP,否则为 FP,如果检测到真实目标没有匹配到对应的预测框则标记为 FN。

3.4　结果分析

3.4.1　改进模型的性能综合分析

如 2.3 节所述,该文尝试将 DRA 模块融合到网络模块的不同位置,并对相应检测结果展开对比。分别在原模型的主干,颈部,检测头中融入 DRA 模块。特殊的对于 backbone 部分,更细化地对比了简单的拼接在尾部或是融入原本的 C3 模块中的结果数据。实验结果如表 1 所示,将 DRA 模块融入主干网络中 C3 模块的最后一层检测效果最佳。YOLOv5 网络中提取特征的关键网络在主干部分,其中隐含着易被网络忽视掉的小目标特征信息,而在加入 DRA 模块后,对这部分的特征信息进行了注意力重构,突出了重要信息,而在网络更深的 Neck 以及 Head 部分,小目标的特征信息被淹没,语义信息较为粗糙,注意力模块难以区分出空间以及通道特征,自然无法很好地对特征进一步加强重构。

表 1 不同位置的注意力机制融合结果对比 %

注意力插入位置	Precision	Recall	mAP@0.5
Yolov5	81.6	79.2	83.1
插入主干 C3 的最后一层	83.7	81.0	87.8
插入主干 C3 残差块之后	81.1	77.8	83.4
插入 Neck 中	82.4	79.3	85.2
插入 Head 中	80.3	80.9	84.8

同时,将文中对 YOLOv5 的注意力及结合方式与其他注意力机制做对比,对比结果如表 2 所示,SE[18]是经典的注意力机制起源,CA[25]是坐标注意力机制,CBAM[20]是经典的空间通道注意力机制,ANG 是一种轻量型的融合注意力机制方法模型。可以看出模型并不适合简单地嵌套所有的注意力机制,当融合 SE 后,模型的漏检率不降反增,说明网络对于深层信息还是没有掌握能力,再看 ANG 模型,轻量化的同时也带来了精度的大量牺牲,而传统的 CA,CBAM 也都基本维持在原精度附近,说明对于网络没有实质性的提升。

表 2 不同注意力机制融合对比结果 %

注意力模型	Precision	Recall	mAP@0.5
Yolov5	81.6	79.2	83.1
SE	81.2	75.8	81.9
CA	83.1	74.3	83.0
CBAM	81.2	79.4	83.3
ANG	81.6	70.3	78.2
Ours	83.7	81.0	87.8

表 3 消融实验结果

改进	SIoU	DRA 融合	mAP@0.5/%	FPS
YOLOv5	×	×	83.1	90.9
1	√	×	85.8	94.4
2	×	√	87.1	78.1
3	√	√	87.8	83.3

为了分析不同的改进策略对于模型最后的检测性能的影响,设计了 4 组消融实验,结果如表 3 所示,其中,“×”代表在网络中未使用的改进策略,“√”代表使用了改进。改进 1 在网络中替换了损失函数,解决了目标框与预测框的角度问题,使模型收敛速度与定位精准度提升;改进 2 在网络主干部分的 C3 模块中融合了 DRA 注意力机制,使得权重中同时包含了通道信息,横向以及纵向空间信息,mAP 提升了 4.0 百分点,FPS 下降了 12.8;改进 3 将两者同时融入网络中,如前文所述,模型在更好地提取特征的同时加快了收敛速度,mAP 最终提升了 4.7 百分点,检测速度则在改进 2 的基础上加速了 5.2,仅与原模型相差 7.6。

同时将消融实验的 mAP@0.5 曲线绘制在同一个坐标系中,如图 5 所示,改进后的模型在迭代次数达到 45 时逐渐趋于稳定。进一步分析 SIoU 改进的数据值曲线,与原始模型的曲线对比,以更高的收敛速度趋于稳定,表明了 SIoU 损失函数的替换使得回归目标框能够以更快的速度,更低的损失,精准地定位到待检测目标。

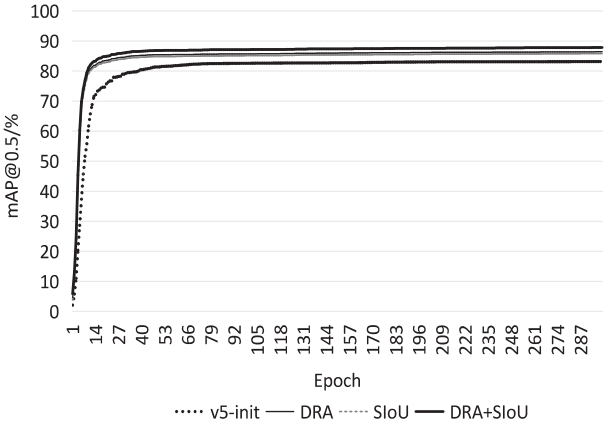


图 5 不同改进策略的 mAP@0.5 对比

3.4.2 模型对比实验

同时将文中模型与其他模型对所有类别检测精度进行对比分析。Faster R-CNN[4]是二阶段检测模型的典型代表,SSD[6]是经典单阶段模型,v3[11]在精度和速度上有较好的均衡性能。v7[26]是当前表现较为出色的检测模型,而 YOLOv5 是文中改进对象。表 4 为所比较模型在所有类别上的检测平均精度对比,在所有 20 类上的检测结果均优于原 v5s 模型,平均精度均值为 87.8%,同时与当前较为优秀的 YOLOv7 模型相比,20 类中有 16 类的结果高于 v7,同时最终的平均精度均值提升 1.4 百分点。

为进一步证实文中算法的有效性和优越性,将文中算法模型与主流模型进行对比。由表 5 中实验结果可得,文中算法模型在保持一定检测速度的情况下,拥有更高的检测精度。与传统的双阶段算法 Faster R-CNN[4]相比具有较大的检测速度优势,平均精度均值提升了 14.6 百分点。与 YOLO 系列算法相比,改进模型比 v3[11],v4[8],v5,v7[26]原始模型的检测精度分别提高了 10.6,15.1,4.7 和 1.7 百分点。对于衡量难检目标以及小目标检测问题的阈值为 0.5 到 0.95 的平均精度均值(mAP@0.5:0.95),对比 v5 提升了 4.5 百分点,对比 v7 提升了 2.8 百分点。而在检测速度方面,文中模型虽比原始模型有所降低,但仍达到 83.3 frame/s,完全可以满足工业场景下的实时检测要求(30 frame/s)。

表 4　VOC 上各类别平均精度结果

Model	Faster R-CNN	SSD	YOLOv3	YOLOv5	YOLOv7	Ours
aero	80.3	79.3	85.5	90.7	92.3	94.5
bike	87.8	83.5	85.3	92	93.1	93.7
bird	75.4	76.6	81.5	81.3	85.2	86.5
boat	64.4	66.5	60.2	72.9	77.2	79.5
bottle	55.9	41.7	55.2	75.3	80.3	85.2
bus	84.3	87.5	85.3	89.5	91.4	94.2
car	86.4	85	88.3	92.7	94.2	94.3
cat	89.4	90.7	91.2	87	87	93.1
chair	60.1	56.5	53.9	67.1	74.6	74.1
cow	79.8	81.7	81.1	88.7	94.1	91.2
table	72.8	77.4	66.7	75.4	84.4	82.9
dog	87.9	87.9	90.6	88.1	85.7	91.9
horse	87.9	87.5	89.4	91.5	91	93.7
mibke	83.3	84.3	83.9	90.6	92	93.2
person	84	74.4	81.2	88.9	91.5	92.9
plant	49.2	48.1	43.2	57.8	64.8	66.7
sheep	75.3	77.6	79.7	85	88.9	87.6
sofa	79.9	81.4	77.9	75.5	83	83.9
train	85	88.5	87.9	87.3	88.8	90.2
tv	73.5	78.4	76.7	84.1	89	88.5
mAP/%	77.1	76.7	77.2	83.1	86.4	87.8

表 5　不同模型的 VOC 数据集测试结果

模型	Precision/%	Recall/%	mAP@0.5/%	mAP@0.5:0.95/%	FPS
Faster R-CNN	73.2	55.2	73.2	44.0	7.0
SSD300	76.8	59.4	76.8	45.6	44.3
YOLOv3	77.2	52.5	77.2	39.8	74
YOLOv4	80.4	62.3	72.7	46.1	54
YOLOv5	79.9	77.9	83.1	59.4	90.9
YOLOv7	84.9	75.1	86.1	60.1	92
Ours	83.7	81.0	87.8	63.9	83.3

4　结束语

YOLO 系列目标检测算法是运用较为广泛的单阶段目标检测算法之一。针对 YOLOv5 对难检目标,包括小目标和遮挡目标等检测精度不高的问题,提出了注意力机制融合的方法,将 DRA 模块与 v5 网络的主干部分进行结合,以增强模型对于一些易漏信息的捕捉能力。同时使用了 SIoU 函数替换原损失函数中负责计算回归参数的 CIoU 损失,提高了收敛速度和回归精度,改善了遮挡等复杂情况下的漏检以及小目标物体识别差的问题。实验结果表明,改进模型的平均精度超越了原 YOLOv5 网络。虽然模型参数量稍有增加,但改进模型的检测速度仍符合工业需求的检测速度。在后期研究中,还可以尝试对于主干网络中的卷积部分进行替换,或是替换特征加强的 Neck 部分,进一步提升模型对于难检目标的检测精度。

参考文献:

[1] XING J, JIA M. A convolutional neural network-based method for workpiece surface defect detection[J]. Measurement, 2021, 176: 109185.

[2] WASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//31st conference on neural information processing systems. Long Beach: MIT Press, 2017: 5998-6008.

[3] 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. Columbus: IEEE, 2014:580-587.
[4] 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.
[5] 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.
[6] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Computer vision - ECCV 2016: 14th European conference. Amsterdam: Springer, 2016: 21-37.
[7] TIAN Z, SHEN C, CHEN H, et al. Fcos: fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. Seoul: IEEE, 2019: 9627-9636.
[8] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. Yolov4: optimal speed and accuracy of object detection[J]. arXiv: 2004.10934, 2020.
[9] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii: IEEE, 2017: 2117-2125.
[10] REDMON J, FARHADI A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Hawaii: IEEE, 2017: 7263-7271.
[11] REDMON J, FARHADI A. Yolov3: an incremental improvement[J]. arXiv: 1804.02767, 2018.
[12] MISRA D. Mish: a self regularized non-monotonic activation function[J]. arXiv: 1908.08681, 2019.
[13] 张 浩, 董锴龙, 孙 欣, 等. 一种基于 IDT-YOLOv5-CBAM 混合算法的密集小目标检测方法: CN115375913A[P]. 2022.
[14] 李永军, 李莎莎, 李孟军, 等. 一种基于改进 YOLOv5 的红外图像行人目标检测方法: CN113688723A[P]. 2021.
[15] 窦其龙, 颜明重, 朱大奇. 基于 YOLO-v5 的星载 SAR 图像海洋小目标检测[J]. 应用科技, 2021, 48(6): 1-7.
[16] 刘闪亮, 吴仁彪, 屈景怡, 等. Bi-PPYOLO tiny: 一种轻量型的机场无人机检测方法[J]. 安全与环境学报, 2023, 23(2): 480-488.
[17] 田 枫, 贾昊鹏, 刘 芳. 改进 YOLOv5 的油田作业现场安全着装小目标检测[J]. 计算机系统应用, 2022, 31(3): 159-168.
[18] 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.
[19] GIRSHICK R. Fast R-CNN[C]//Proceedings of the IEEE international conference on computer vision. Santiago: IEEE, 2015: 1440-1448.
[20] 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: 3-19.
[21] CAO Y, XU J, LIN S, et al. GCNet: non-local networks meet squeeze-excitation networks and beyond[C]//2019 IEEE/CVF international conference on computer vision workshop (ICCVW). Seoul: IEEE, 2019: 1971-1980.
[22] VZHENG Z, WANG P, REN D, et al. Enhancing geometric factors in model learning and inference for object detection and instance segmentation[J]. IEEE Transactions on Cybernetics, 2021, 52(8): 8574-8586.
[23] GEVORGYAN Z. SIoU loss: more powerful learning for bounding box regression[J]. arXiv: 2205.12740, 2022.
[24] KINGMA D P, BA J. Adam: a method for stochastic optimization[J]. arXiv: 1412.6980, 2014.
[25] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. [s.l.]: IEEE, 2021: 13713-13722.
[26] 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[J]. arXiv: 2207.02696, 2022.