代涛, 黎青松, 邓鹏毅. 基于改进 YOLOv5 的小目标检测算法研究[J]. 智能计算机与应用,2025,15(9):132-138. DOI:10.20169/j.issn.2095-2163.250921

基于改进 YOLOv5 的小目标检测算法研究

代 涛, 黎青松, 邓鹏毅
(西华大学 汽车与交通学院, 成都 610039)

摘 要: 目前关于大、中目标的检测已经取得较为可观研究成果,而小目标因其像素少、携带信息少,在下采样过程中信息易丢失等原因,导致小目标的检测结果欠佳。针对小目标检测的问题,提出了一个以 YOLOv5 为基线的改进算法。首先,在该算法中添加了一个检测头,实现更细粒度的目标检测;然后,添加 BiFormer 双层路由注意力机制来增加小目标的重要性;针对 IoU 计算时对小目标检测的不良影响,使用 Wasserstein 距离来度量 BBox 的相似性,从而代替标准 IoU 的计算。实验证明,改进后的算法比原算法的 mAP@0.5 提高了 0.045 2,验证了改进算法对提高小目标检测效果的有效性。

关键词: YOLOv5; 小目标检测头; BiFormer; Wasserstein 距离
中图分类号: TP391.4 **文献标志码**: A **文章编号**: 2095-2163(2025)09-0132-07

Research on small target detection algorithm based on improved YOLOv5

DAI Tao, LI Qingsong, DENG Pengyi
(School of Automobile and Transportation, Xihua University, Chengdu 610039, China)

Abstract: At present, there have been many research results on the detection of large and medium targets, and considerable fruits have been achieved. However, small targets have poor detection results due to their few pixels, carrying less information, and easy information loss during downsampling. An improved algorithm based on YOLOv5 is proposed to address the issue of small object detection. Firstly, a detection head is added to the algorithm to achieve finer grained object detection. Then, BiFormer dual layer routing attention mechanism is added to increase the importance of small targets. To address the negative impact of IoU calculation on small object detection, Wasserstein distance is used to measure the similarity of BBox instead of standard IoU calculation. Experimental results have shown that compared with the original algorithm, mAP@0.5 of the improved algorithm is increased by 0.045 2, verifying the effectiveness of the improved algorithm in improving the performance of small object detection.

Key words: YOLOv5; small target detection head; BiFormer; Wasserstein distance

0 引 言

随着大数据时代的到来,卷积神经网络(Convolution Neural Network,CNN)得到快速发展,基于卷积神经网络的目标检测算法也日益成熟[1]。小目标检测在航空航天、遥感图像等领域应用广泛,所以小目标检测对多个领域的研究推进都具有重要意义。

然而现有目标检测算法对大、中目标的检测效果较好,小目标因为其分辨率低、在图像中覆盖面积较少、特征表达不充分等弊端,导致对于小目标的检测仍未取得令人满意效果。因此有必要对小目标检测展开更深入的研究。

在不同场景下小目标的定义标准不尽相同,但根据现有的定义方式在学术上主要分为 2 类:

(1)根据绝对尺寸定义。在 MS COCO 数据集[2]中将尺寸小于 32×32 像素的目标定义为小目标。

(2)根据相对尺寸定义。在图像中占比低于 1%的目标,被定义为小目标。

目前,基于深度学习的目标检测算法主要分为 2 类,即双阶段目标检测算法(Two Stage)和单阶段目标检测算法(One Stage)[3]。其中,双阶段目标检测算法主要以 R-CNN[4]、Fast R-CNN[5]以及 Faster

基金项目: 四川省自然科学基金(2021YFG0070)。
作者简介: 代 涛(1996—),男,硕士研究生,主要研究方向:目标检测算法研究; 邓鹏毅(1972—),男,副教授、高级工程师,硕士生导师,主要研究方向:目标检测算法研究。
通信作者: 黎青松(1973—),男,博士,教授、高级工程师,硕士生导师,主要研究方向:目标检测算法研究。Email:liqs73@163.com。
收稿日期: 2024-01-02

R-CNN[6]为代表，这类算法将检测问题划分为 2 个阶段。第 1 个阶段产生目标候选区域，初步确定目标位置；第 2 个阶段对目标候选区域进行目标分类和位置精修。双阶段目标检测算法的优势在于检测精度较高，缺点在于检测速度较慢。单阶段目标检测算法主要以 SSD[7]和 YOLO[8-11]系列为代表，单阶段目标检测算法省去了候选框生成步骤，而是把图像中所有的位置都视作潜在的目标进行处理。相对于双阶段目标检测算法，单阶段目标检测算法检测速度更快，但是检测精度较低。

此外，因为 SSD 对于目标检测表达能力存在欠缺，而且也未能考虑卷积网络不同层特征图的融合，而 YOLO 目标检测算法相对于 SSD 而言，一直在迭代更新，性能也在不断提升。综上所述，本文将以单阶段目标检测算法 YOLO 为基线进行研究。

1　YOLOv5 目标检测算法改进

1.1　YOLOv5 目标检测算法介绍

YOLOv5 是 YOLO 系列算法的第 5 版，由 Ultralytics 于 2020 年 5 月提出，是具有优秀性能的单阶段目标检测算法。YOLOv5 有 YOLOv5s、YOLOv5m、YOLOv5l、YOLOv5x 四种网络结构。其中，YOLOv5s 网络是 YOLOv5 系列中深度最小、特征图的宽度最小的网络，后面 3 种都是在此基础上不断加深、不断加宽，检测速度也是越来越慢的。综上可得，本文选取 YOLOv5l 作为基础模型，兼顾检测精度和检测速度，通过调整网络结构，来优化算法性能，提升目标检测效果。

研究可知，YOLOv5 主要包含 Input、Backbone、Neck、Prediction 四部分。

总地来说，Input 是输入图像，YOLOv5 的输入端使用了 Mosaic 的数据增强方式，在增强模型鲁棒性的同时减少了对 GPU 的占用；Backbone 是特征提取网络，主要用来提取输入图像的特征。YOLOv5 采用 CSPDarkNet53 网络作为 Backbone；Neck 主要是融合不同层次的图特征，随着特征提取的不断深入，图像的某些局部信息会消失，利用 Neck 网络融合不同网络层次的特征图，可以获取图像更丰富的特征信息，再将这些处理后的特征输入 Prediction 层，YOLOv5 采用 PANet[12]进行特征融合；在 Prediction 层共有 3 个检测头，也就是 P3、P4、P5，分别检测不同大小的目标。

1.2　小目标检测头的添加

原版本的 YOLOv5 主干网络一共进行了 5 次下采样。每经历一次下采样，特征图的长宽就会变为原来的一半。如果输入图像是 640×640×3 大小的图片，经过特征提取网络 5 次下采样后，则会输出大小为 20×20 的特征图，再经过底层与高层的特征融合，可以进行 32×32 大小的目标检测，这就是 P5 检测头；再经过一次上采样后，特征图就会变成 40×40 大小，此时和特征提取网络中第 4 次下采样操作得到的 40×40 大小的特征图进行融合，就可以检测 16×16 大小的目标，这就是 P4 检测头；40×40 大小的特征图再经过一次上采样后，变成 80×80 大小的特征图，将其和特征提取网络中第 3 次下采样操作得到的 80×80 的特征图进行融合，可以检测 8×8 大小的目标，这就是 P3 检测头。

图片在经过特征网络多次下采样后，特征图中的目标信息变得非常稀少，原有的 3 个检测层已无法满足对微小目标的检测需要，容易产生漏检现象，因此本文在 P3 检测头后面再加一个检测头，形成一个新的检测头 P2[13]，用来检测大小为 4×4 的目标，而且该检测头的高分率为预测特征图中保留了更多关于小目标的位置信息和更丰富的细节特征，可以更好地获取到小目标的位置信息。改进后的 YOLOv5 算法结构如图 1 所示。

1.3　注意力机制的添加

注意力机制起源于自然语言处理，是为了更好地联系上下文的关系，方便理解全篇内容，后来工程师将其应用于 CNN 网络，大大提高了网络对特征的提取能力，取得了非常不错的效果。随着时间的发展、任务的不同，出现了各种各样的注意力机制，如 CBAM[14]、SENet[15]等等，而在本次研究中采用的是 BiFormer[16]双层路由注意力机制。

BiFormer 是 2023 年提出的一个注意力机制，设计的目的是为了应用于密集目标检测，并减少计算量，减轻计算机的负担，然而其对小目标检测效果也有着不小的提升。

BiFormer 对图像进行 self-attention[17]，是基于稀疏采样而不是下采样。相比于下采样，稀疏采样不仅可以保留细粒度的细节信息，而且也可以节省计算量。

BiFormer 原理如图 2 所示，BiFormer 的工作主要包含以下 3 个步骤：

(1)将特征图划分区域并进行线性映射。首先将输入的特征图划分为 $S\times S$ 个非重叠区域，使每个区域包含 $\frac{HW}{S^2}$ 个特征向量。此后，通过线性映射推导出特征图的 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$，得到的线性映射为：

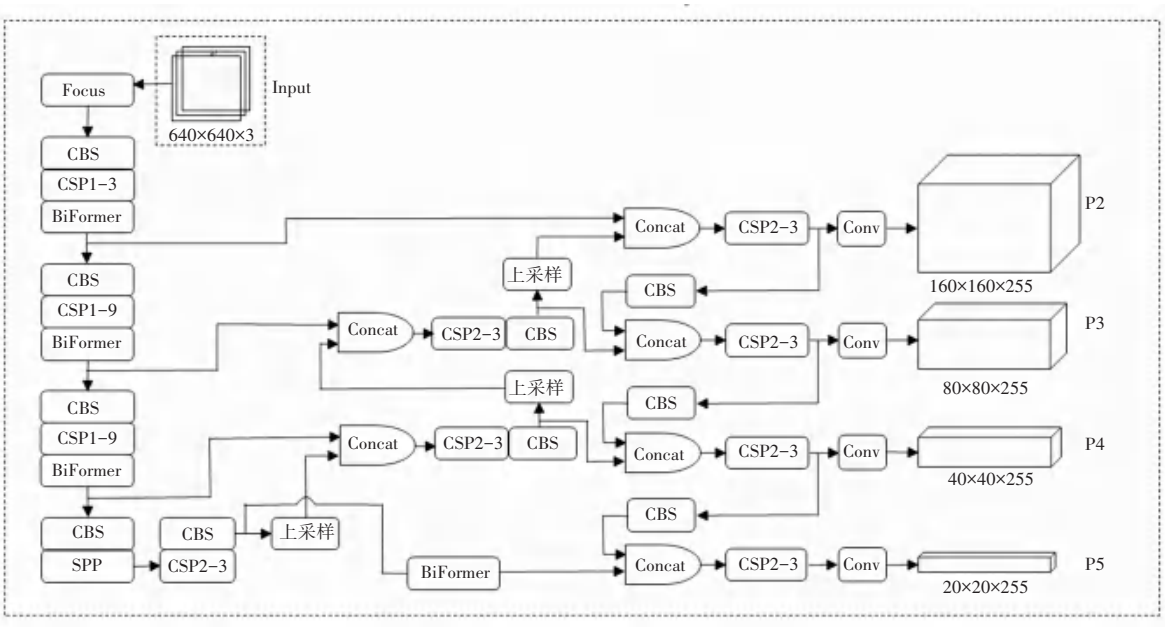


图 1 改进后的 YOLOv5 算法结构图

Fig. 1 Structure diagram of improved YOLOv5 algorithm

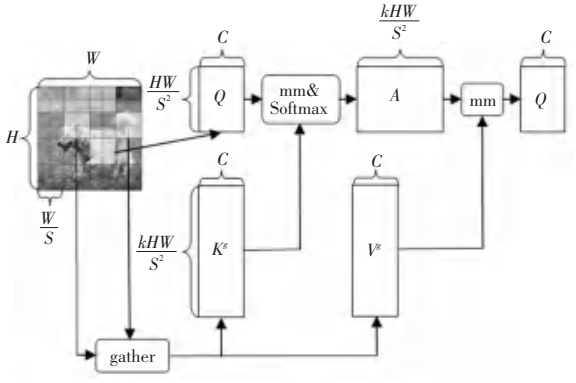


图 2 BiFormer 原理图

Fig. 2 Schematic diagram of BiFormer

$$\boldsymbol{Q}=\boldsymbol{X}^{r}\boldsymbol{W}^{q},\ \boldsymbol{K}=\boldsymbol{X}^{r}\boldsymbol{W}^{k},\ \boldsymbol{V}=\boldsymbol{X}^{r}\boldsymbol{W}^{v} \tag{1}$$

(2)对图片进行粗粒度的注意力处理,并提取关联度比较高的区域。使用平均池化将得到的查询与键导出每个区域的 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$,通过矩阵乘法推导出区域到区域之间亲和度的邻接矩阵:

$$\boldsymbol{A}^{r}=\boldsymbol{Q}^{r}(\boldsymbol{K}^{r})^{\mathrm{T}} \tag{2}$$

$\boldsymbol{I}^{r}$ 的第 i 行包含了第 i 个区域中最相关区域的 k 个指数。然后通过邻接矩阵推导出路由索引矩阵:

$$\boldsymbol{I}^{r}=\mathrm{topkIndex}(\boldsymbol{A}^{r}) \tag{3}$$

最后根据路由索引矩阵就可以寻找到关联比较大的区域。

(3)在关联度比较大的区域内,再进行细粒度的注意力处理,以提取重要信息。在筛选出关联度比较大的粗粒度区域后,应用细粒度的 token-to-token[18-19] 的注意力关注,应用到收集到的键值对上:

$$\boldsymbol{O}=\mathrm{Attention}(\boldsymbol{Q},K^{g},V^{g})+\mathrm{LCE}(\boldsymbol{V}) \tag{4}$$

其中,函数 LCE($\boldsymbol{V}$) 是通过深度卷积进行参数化。

简单梳理下,假设输入一张特征图,通过线性映射获得 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$;随后,通过邻接矩阵构建有向图找到不同键值对的参与关系,可以理解为每个给定区域应该参与的区域;最后,有了区域到区域路由索引矩阵,就可以对连接比较密切的区域应用细粒度的 token-to-token 注意力了。

1.4 NWD 度量

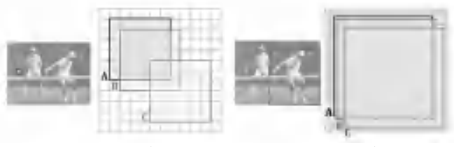
IoU 是在特定数据集中检测相应物体准确度的一种量化标准。

在目标检测中,预测的边框和真实的边框的交集和并集的比值就是 IoU。

IoU 对不同尺度物体的敏感性差异很大,在基于 Anchor 的检测器中使用时,会严重降低检测器的检测性能。具体来说,对于微小物体,细微的位置偏差将使 IoU 值出现显著下降,从而导致标签分配不准确。然而,对于大、中目标而言,相同的位置偏差只会导致 IoU 略有变化。

更进一步地,IoU 对微小和正常尺度物体的敏感性分析如图 3 所示。图 3 中,每个网格表示一个像素,方框 A 表示地面真实边界框,方框 B、C 分别表示对角线偏差为 1 像素和 4 像素的预测边界框。对于 6×6 像素的小目标而言,轻微的位置偏差会使 IoU 值下降明显(从 0.53 下降到 0.06),导致标签分配不准确。然而,对于 36×36 像素的正常目标,相同的位

置偏差，IoU 只是略有变化（从 0.90 到 0.65）。



（a）小目标　　（b）大、中目标

图 3　IoU 对微小和正常尺度物体的敏感性分析

Fig. 3　The sensitivity analysis of IoU on tiny and normal scale objects

针对此种情况，本文使用了一种新的度量标准，称为 Normalized Wasserstein Distance[20]，简称 NWD，其核心思想是使用 Wasserstein 距离代替 IoU 作为微小物体检测的评估指标。NWD 的工作主要包含以下 2 个步骤：

（1）边界框的高斯分布建模。对于微小的物体，其边界框中往往会有一些背景像素，因为大多数真实物体不是严格的矩形。而在这些边界框中，前景像素和背景像素分别集中在边界框的中心和边界上。

为了更好地描述边界框中不同像素的权重，边界框可以建模为二维（2D）高斯分布，其中边界框的中心像素具有最高的权重，像素的重要性从中心到边界递减。

对于水平边界框 $R=(cx,cy,w,h)$，其中，(cx,cy)，w 和 h 分别表示中心坐标、宽度和高度。由此推得的内接椭圆方程可以表示为：

$$\frac{(x-\mu_x)^2}{\sigma_x^2}+\frac{(y-\mu_y)^2}{\sigma_y^2}=1 \tag{5}$$

其中，(μ_x,μ_y) 表示椭圆的中心坐标，σ_x 和 σ_y 分别表示沿 x 和 y 轴的半轴长度。因此，$\mu_x=cx$，$\mu_y=cy$，$\sigma_x=\frac{w}{2}$，$\sigma_y=\frac{h}{2}$。二维高斯分布的概率密度函数的数学定义为：

$$f(x/\mu,\boldsymbol{\Sigma})=\frac{\exp\left(-\frac{1}{2}(x-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(x-\mu)\right)}{2\pi\,|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \tag{6}$$

其中，x、μ 和 $\boldsymbol{\Sigma}$ 分别表示高斯分布的坐标（x，y）、均值向量和协方差矩阵。还要提及的是，需将 $\boldsymbol{\Sigma}$ 与累加求和符号进行区分。

当满足 $(X-\mu)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(X-\mu)=1$ 时，式（5）中的椭圆将会是二维高斯分布的密度等值线。因此，水平边界框 $R=(cx,cy,w,h)$ 可以建模为二维高斯分布 $N(\mu,\boldsymbol{\Sigma})$，分析推得的数学公式如下：

$$\mu=\begin{bmatrix}cx\\cy\end{bmatrix},\ \boldsymbol{\Sigma}=\begin{bmatrix}\frac{w^2}{4} & 0\\ 0 & \frac{h^2}{4}\end{bmatrix} \tag{7}$$

这样一来，边界框 A 和 B 之间的相似度就可以转换为 2 个高斯分布之间的分布距离。

（2）归一化高斯 Wasserstein 距离。对于 2 个二维高斯分布 $\mu_1=N(m_1,\boldsymbol{\Sigma}_1)$ 和 $\mu_2=N(m_2,\boldsymbol{\Sigma}_2)$，$\mu_1$ 和 μ_2 之间的二阶 Wasserstein 距离定义为：

$$W_2^2(\mu_1,\ \mu_2)=\|m_1-m_2\|_2^2+Tr(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2-2(\boldsymbol{\Sigma}_2^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2^{1/2})^{1/2}) \tag{8}$$

研究中，式（7）可以简化为：

$$W_2^2(\mu_1,\mu_2)=\|m_1-m_2\|_2^2+\|\boldsymbol{\Sigma}_1^{1/2}-\boldsymbol{\Sigma}_2^{1/2}\|_F^2 \tag{9}$$

其中，$\|\cdot\|_F^2$ 表示 Frobenius 范数。

对于从边界框 $A=(cx_a,cy_a,w_a,h_a)$ 和 $B=(cx_b,cy_b,w_b,h_b)$ 建模的高斯分布 N_a 和 N_b，式（8）可进一步简化为：

$$W_2^2(N_a,N_b)=\left\|\left[cx_a,cy_a,\frac{w_a}{2},\frac{h_a}{2}\right]^{\mathrm{T}}-\left[cx_b,cy_b,\frac{w_b}{2},\frac{h_b}{2}\right]^{\mathrm{T}}\right\|_2^2 \tag{10}$$

但是，$W_2^2(N_a,N_b)$ 是距离度量，不能直接用作相似度度量（IoU 的值为 0 ~ 1）。因此，需要对 $W_2^2(N_a,N_b)$ 进行归一化，获得归一化 Wasserstein 距离（NWD）的新度量：

$$\mathrm{NWD}(N_a,N_b)=\exp\left(-\frac{\sqrt{W_2^2(N_a,N_b)}}{C}\right) \tag{11}$$

将 NWD 度量设计为损失函数代替 IoU 损失函数：

$$L_{\mathrm{NWD}}=1-\mathrm{NWD}(N_a,N_b) \tag{12}$$

其中，N_a 表示预测框 A 的高斯分布模型；N_b 表示真实框 B 的高斯分布模型。

NWD 的优点在于预测框和真实框即使没有重叠或重叠可以忽略不计，也可以测量分布的相似性。而且，NWD 对不同尺度的物体不敏感，因此更适合用于微小物体之间的相似性测量。

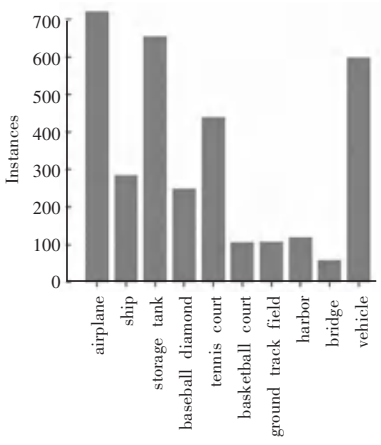
2　实验与结果分析

2.1　实验配置

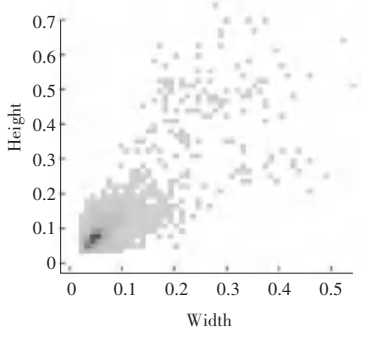
本文实验的计算机配置如下：Win11 操作系统；深度学习框架 Pytorch，版本是 1.12；Python 版本是 3.9.12；Cuda 版本是 12.2，Cudnn 版本是 8.9；CPU 是 i7-12700H；GPU 是 NVIDIA RXT 3050Ti。

2.2　数据集

本文使用的是无人机航拍小目标数据集 VisDrone2019 数据集[21]，具体如图 4 所示。



(a) 类别及数量



(b) 目标大小及占比

图 4 VisDrone2019 数据集

Fig. 4 VisDrone2019 dataset

由图 4 可知,VisDrone2019 数据集主要包括 10 个类别,而且,绝大部分目标均为小目标,大小只有几十个像素。所以,该数据集足以保证对小目标检测算法的训练和验证。

2.3 评价指标

这里对主要采用的评价指标做阐释分析如下。

(1)精确度(Precision, P)和召回率(Recall, R)。数学定义公式为:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{13}$$

$$R = \frac{TP}{TP + FN} \times 100\% \tag{14}$$

其中,TP 表示实际为正例且被分类器划分为正例的样本数;FP 表示实际为负例但被分类器划分为正例的样本数;FN 表示实际为正例但被分类器划分为负例的样本数。

(2)平均精准度(Average Precision, AP)和(mean Average Precision, mAP)。AP 即是以 Precision 与 Recall 分别作为纵坐标和横坐标时曲线所围成的面积的大小。mAP 表示所有类别的 AP 值取平均,计算公式如下:

$$AP = \int_0^1 P(R)\,dR \tag{15}$$

$$mAP = \frac{\sum_{f=1}^{n} AP_f}{n} \tag{16}$$

2.4 消融实验及结果分析

本文以 YOLOv5l 为基线算法,基于 VisDrone2019 数据集,通过消融实验探究每个新增或改进模块对于整体模型的提升效果。

本文的训练方案如下:batch size 为 4;epoch 为 200;初始学习率为 0.01;周期学习率为 0.2;并使用 YOLOv5l 作为预训练权重,分别对各个模块进行消融实验。

消融实验结果见表 1。表 1 中,mAP@0.5 表示 IoU 为 0.5 时所有类别的 mAP,mAP@0.50:0.95 表示 IoU 以 0.05 的步长从 0.50 到 0.95 时的 mAP。由表 1 可知,原 YOLOv5l 算法在 VisDrone2019 数据集上训练后的 mAP@0.5 为 0.376 2,mAP@0.50∶0.95 为 0.211 9,Precision 为 0.492 3,Recall 为 0.385 6。此后分别添加小目标检测头 P2、BiFormer 和 NWD 模块进行消融实验。

表 1 消融实验结果

Table 1 Results of ablation experiment

算法	mAP@0.5	mAP@0.50∶0.95	Precision	Recall
YOLOv5l	0.376 2	0.211 9	0.492 3	0.385 6
YOLOv5l+P2	0.399 6	0.225 0	0.514 5	0.409 9
YOLOv5l+BiFormer	0.396 6	0.222 1	0.497 3	0.413 0
YOLOv5l+NWD	0.381 5	0.212 0	0.507 2	0.391 6
YOLOv5l+P2+BiFormer+NWD	0.421 4	0.244 3	0.549 8	0.431 4

YOLOv5 首先添加小目标检测头。添加小目标检测头后算法的 mAP@0.5 增加了 0.023 4，是所有模块当中提升最大的；添加 BiFormer 注意力机制后，算法的 mAP@0.5 增加了 0.020 4，提升的效果稍逊于小目标检测头；而用 NWD 替换 IoU，以减低 IoU 计算时对小目标的不良影响，算法的 mAP@0.5 仅仅提升了 0.005 3，是所有改进当中提升最小的，可见 NWD 对小目标检测提升的效果有限。

然后把所有改进集成到 YOLOv5l 算法上，从表 1 可以看出，改进后的算法相较于原算法 mAP@0.5 提升了 0.045 2，达到了 0.421 4，而 mAP@0.50：0.95、Precision 以及 Recall 也有所提升，分别提升了 0.032 4、0.057 5 以及 0.045 8。

最后对原算法和改进后的算法进行可视化，本文选择了 3 个场景，分别是俯拍车流、密集人群以及多尺度场景。可视化效果如图 5 所示。从图 5 可以看出，在俯拍车流场景下，改进算法可以检测到远处更小的目标；在密集人群中，对于原算法没有检测到的遮挡的人群，改进算法也检测到了；而在多尺度场景中，对于原算法漏检的摩托车以及远处的小目标，改进算法都检测到了。综上分析，说明改进算法可以有效缓解因为目标像素小、密集分布且相互遮挡、在多尺度场景下引发的漏检情况，表明本文改进的方法是有效的。

(a) 俯拍车流

(b) 密集人群

(c) 多尺度场景

图 5　原算法和改进后的算法可视化

Fig. 5　Visualization of the original algorithm and the improved algorithm

3　结束语

本文针对小目标检测过程中，小目标由于像素较少、携带信息少，在下采样过程中信息易丢失等原因、导致小目标检测效果并不理想的情况，提出了一种基于 YOLOv5 的改进算法。具体步骤包括：

(1)添加小目标检测头，用于对小目标的检测。

(2)添加 BiFormer 双层路由注意力机制，从粗粒度和细粒度两个方面进行注意力处理，提高小目标的重要性。

(3)针对 IoU 计算时，对小目标的不利影响，使用 NWD 来代替 IoU。

经过实验验证，改进后的算法在 mAP@0.5、mAP@0.50:0.95、Precision 以及 Recall 上均有所提升，实际检测的效果也比较理想。综上所述，改进后的算法可以有效满足小目标的检测要求。

参考文献

[1] JIAO Licheng, ZHANG Fan, LIU Fang, et al. A survey of deep learning based object detection[J]. IEEE Access, 2019, 7: 128837-128868.

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

[3] 段仲静, 李少波, 胡建军,等. 深度学习目标检测方法及其主流框架综述[J]. 激光与光电子学进展, 2020, 57(12): 59-74.

[4] 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, NJ: IEEE, 2014: 580-587.

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

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

[7] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector[C]//Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016: 21-37.

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

[9] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2017: 7263-7271.

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

[11] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4:

Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv, 2004.10934, 2020.
[12] LIU Shu, QI Lu, QIN Haifang, et al. Path aggregation network for instance segmentation [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8759-8768.
[13] 杨慧剑, 孟亮. 基于改进的 YOLOv5 的航拍图像中小目标检测算法[J]. 计算机工程与科学, 2023, 45(6): 1063-1070.
[14] 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, 2018: 3-19.
[15] HU Jie, SHEN Li, ALBANIE S, et al. Squeeze and excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7132-7141.
[16] ZHU Lei, WANG Xinjiang, KE Zhanghan, et al. BiFormer: Vision transformer with bi-level routing attention [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 10323-10333.
[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]// Advances in Neural Information Processing Systems. Long Beach, USA: NIPS Foundation, 2017: 5998-6008.
[18] YUAN Li, CHEN Yunpeng, WANG Tao, et al. Tokens-to-token vit: Training vision transformers from scratch on imagenet [C]//Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE, 2021: 558-567.
[19] QUAN Yu, ZHANG Dong, ZHANG Liyan, et al. Centralized feature pyramid for object detection [J]. IEEE Transactions on Image Processing, 2023, 32: 4341-4354.
[20] WANG Jinwang, XU Chang, YANG Wen, et al. A normalized Gaussian Wasserstein distance for tiny object detection [J]. arXiv preprint arXiv, 2110.13389, 2022.
[21] ZHU Pengfei, WEN Longyin, DU Dawei, et al. Vision meets drones: Past, present and future [J]. arXiv preprint arXiv, 2001.06303, 2020.