计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(12):242−250 [doi: 10.15888/j.cnki.csa.008835]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于改进 YOLOv5 的小目标检测①

黎学飞[1], 童　晶[1], 陈正鸣[1], 包　勇[2], 倪佳佳[1]

[1](河海大学 物联网工程学院, 常州 213022)
[2](江苏医像信息技术有限公司, 常州 213022)
通信作者: 陈正鸣, E-mail: zmchen@hhuc.edu.cn

摘　要: 本文针对图像中小目标难以检测的问题, 提出了一种基于 YOLOv5 的改进模型. 在主干网络中, 加入 CBAM 注意力模块增强网络特征提取能力; 在颈部网络部分, 使用 BiFPN 结构替换 PANet 结构, 强化底层特征利用; 在检测头部分, 增加高分辨率检测头, 改善对于微小目标的检测能力. 本文算法在人脸瑕疵数据集和无人机数据集 VisDrone2019 两份数据集上均进行了多次对比实验, 结果表明本文算法可以有效地检测小目标.

关键词: 小目标检测; 注意力机制; 特征融合; YOLOv5; BiFPN

引用格式: 黎学飞,童晶,陈正鸣,包勇,倪佳佳.基于改进 YOLOv5 的小目标检测.计算机系统应用,2022,31(12):242–250. http://www.c-s-a.org.cn/1003-3254/8835.html

Small Target Detection Based on Improved YOLOv5

LI Xue-Fei[1], TONG Jing[1], CHEN Zheng-Ming[1], BAO Yong[2], NI Jia-Jia[1]

[1](College of Internet of Things Engineering, Hohai University, Changzhou 213022, China)
[2](Jiangsu Medical Image Information Technology Co. Ltd., Changzhou 213022, China)

Abstract: In this study, an improved model based on you only look once version 5 (YOLOv5) is proposed to solve the problem of difficult detection of small targets in images. In the backbone network, a convolutional block attention module (CBAM) is added to enhance the network feature extraction ability. As for the neck network, the bi-directional feature pyramid network (BiFPN) structure is used to replace the path aggregation network (PANet) structure and thereby strengthen the utilization of low-level features. Regarding the detection head, a high-resolution detection head is added to improve the ability of small target detection. A number of comparative experiments are conducted, respectively, on a facial blemish dataset and an unmanned aerial vehicle (UAV) dataset VisDrone2019. The results show that the proposed algorithm can effectively detect small targets.

Key words: small target detection; Attention mechanism; feature fusion; YOLOv5; bi-directional feature pyramid network (BiFPN)

目标检测是计算机视觉的重要研究方向, 也是众多复杂视觉任务的基础, 被广泛应用于工业、农业等领域[1]. 尽管深度学习技术的出现使得目标检测取得了较大的突破, 但是现有的方法依然很难较好的检测小目标. 小目标尺寸小、特征不明显, 在检测中误检率、漏检率一般均较高, 因此提升小目标检测性能仍然是一个具有挑战性的研究方向[2].

2012 年, AlexNet 在 ImageNet 图像识别比赛中夺冠以后, 神经网络迅速发展[3]. 凭借其强大的特征表达能力和建模能力, 神经网络在图像领域快速发展并得到广泛应用, 目标检测也由此得到了快速发展. 目前, 基于深度学习的目标检测主要分为两类: 1) 两阶段检测器, 如基于区域的 RCNN[4] 及其变体[5−7]. 2) 一阶段检测器, 如 SSD[8] 和 YOLO 系列[9] 及其变体[10,11]. 两阶

① 基金项目: 国家重点研发计划 (2020YFB1708900); 常州市重点研发计划 (CE20210045); 江苏省重点研发计划 (BE2020762)
收稿时间: 2022-03-23; 修改时间: 2022-04-21; 采用时间: 2022-05-11; csa 在线出版时间: 2022-08-12

段检测器首先需要生成可能包含目标的候选框, 然后使用区域分类器预测; 一阶段检测器直接对特征图的各位置上的目标进行分类预测, 其更省时且具有更大的实用性. 在 YOLO 系列中, YOLOv3 最为经典, 其在 YOLOv2 上增加一个特征金字塔, 使得其自上而下多级预测的结构增强了网络对于小目标的检测能力.

YOLOv5 是继 YOLOv3 之后被广泛运用于工业检测的算法, 能够在保持较高精确率的同时满足实时性要求, 并且可以根据不同的工作环境和检测任务选择不同的检测型号. Song 等人[12] 将 YOLOv5 应用于道路密集车辆实时检测. Sruthi 等人[13] 将 YOLOv5 搭载与搜救无人机上, 用于搜救过程中的人体检测. Dong 等人[14] 将 YOLOv5 结合注意力机制, 提出了一种用于检测肺结节的方法. YOLOv5 利用深度残差网络提取目标特征, 并利用 PANet 结构完成了多尺度预测, 但是 YOLOv5 提取特征获得最大特征图时依然进行了 3 次下采样, 目标特征信息丢失过多, 对于微小目标的检测并不理想. 因此, 本文对 YOLOv5 进行改进, 提出了 BCF-YOLOv5 检测模型, 用于小目标检测. 主要贡献如下: 1) 针对小目标对象提出了一种新的检测模型. 2) 基于注意力机制, 通过改进主干网络增强模型对于目标特征的提取. 3) 基于双向交叉连接和加权融合的思想, 引入额外特征分支, 强化底层特征利用. 4) 针对目标尺寸较小的特点, 在检测头部分增加高分辨率检测分支, 增强模型对于微小目标的检测能力.

1 BCF-YOLOv5 模型

YOLOv5 由 Ultralytics 于 2020 年 5 月提出[15], 是性能非常优秀的一阶段检测器, 所以选择它作为我们的基础模型. YOLOv5 有 YOLOv5s、YOLOv5m、YOLOv5l 、YOLOv5x 四个模型. 通过 depth_multiple 和 width_multiple 两个参数控制模型的复杂度.其中 YOLOv5s 模型最小, 检测速度最快, 但是检测精度最低; YOLOv5x 模型最大, 检测精度最高, 但是检测速度最慢. 一般来说, YOLOv5 采用 CSPDarknet53 框架加一个 SPP 层作为骨干, 搭配 PANet 颈部和 YOLO 检测头. 尽管 YOLOv5x 模型的计算成本高于其他 3 个模型, 但是为获得最好的检测性能, 本文依然选择 YOLOv5x 作为基础模型. 针对小目标尺寸小、特征不明显的特点, 我们修改了最初的 YOLOv5, 提出了 BCF-YOLOv5 模型, 使其更加适用于微小目标的检测, 其网络结构如图 1 所示.

BCF-YOLOv5 模型在结构上分为 4 部分: 输入、主干网络、颈部网络和检测头部分. 输入部分将训练数据进行 Mosaic、Mixup 等数据增强, 然后对图片进行自适应缩放以及自适应锚框计算. 主干网络包括 Focus、CBAM (convolutional block attention module)[16]、CSP 和 SPP 结构, Focus 通过自我复制然后进行切片操作, 加快网络推理. CBAM 注意力模块提高主干网络对小目标的特征提取能力. CSP 残差结构优化梯度信息传递, 减少推理计算量. SPP 空间金字塔结构分别采用直连和 5、9、13 的最大池化, 然后进行 concat 融合, 提高感受野. 颈部网络部分采用 BiFPN 结构, 以双向交叉连接方式进行特征融合, 得到一系列分层特征表示, 并将这些特征表示传递给检测头部分. 检测头输出依然采用 CIoU 作为 Bounding box 的损失函数, 并提供 4 种检测尺度 (20×20、40×40、80×80、160×160). 最后通过 NMS 对目标检测框进行非极大值抑制获得最优预测框.

1.1 小目标检测头

原始 YOLOv5 模型主干网络一共进行 5 次下采样, 得到 5 层特征表达 ($P1$、$P2$、$P3$、$P4$ 和 $P5$), 其中 Pi 表示分辨率为原始图片的 $1/2^i$, 尽管在颈部网络中通过自上而下和自下而上的聚合路径实现了多尺度特征融合, 但是并不影响特征图的尺度, 最后检测头部分在通过 $P3$、$P4$ 和 $P5$ 这 3 级特征图引出的检测头上进行目标的检测, 其特征图尺度分别为 80×80、40×40、20×20. 为方便表达, 通过 Pi 层特征图引出的检测头, 以下简称为 Pi 层检测头. 在小目标检测任务中, 往往存在非常小的待检目标. 在本文自建的人脸瑕疵数据中, 包含较多的诸如点状色斑、小黑痣等微小瑕疵, 其尺度往往小于 6×6 像素. 在 VisDrone2019[17] 公开无人机数据集中, 甚至含有较多小于 3×3 像素的目标. 这样的目标在经过多次下采样之后, 其大部分特征信息已经丢失, 尽管通过具有较高分辨率的 $P3$ 层检测头依然难以检测到.

为了实现上述微小目标同样可以达到较好的检测效果, 我们在 YOLOv5 模型上通过 $P2$ 层特征引出了新的检测头. 结构如图 2 所示. $P2$ 层检测头分辨率为 160×160 像素, 相当于在主干网络中只进行了 2 次下采样操作, 含有目标更为丰富的底层特征信息. 颈部网络中自上而下和自下而上得到的两个 $P2$ 层特征与主干网络中的同尺度特征通过 concat 形式进行特征融合,

输出的特征为 3 个输入特征的融合结果, 这样使得 *P*2 层检测头应对微小目标时, 能够快速有效的检测. *P*2 层检测头加上原始的 3 个检测头, 可以有效缓解尺度方差所带来的负面影响. 增加的检测头是针对底层特征的, 是通过低水平、高分辨率的特征图生成的, 该检测头对微小目标更加敏感. 尽管添加这个检测头增加了模型的计算量和内存开销, 但是对于微小目标的检测能力有着不小的提升.

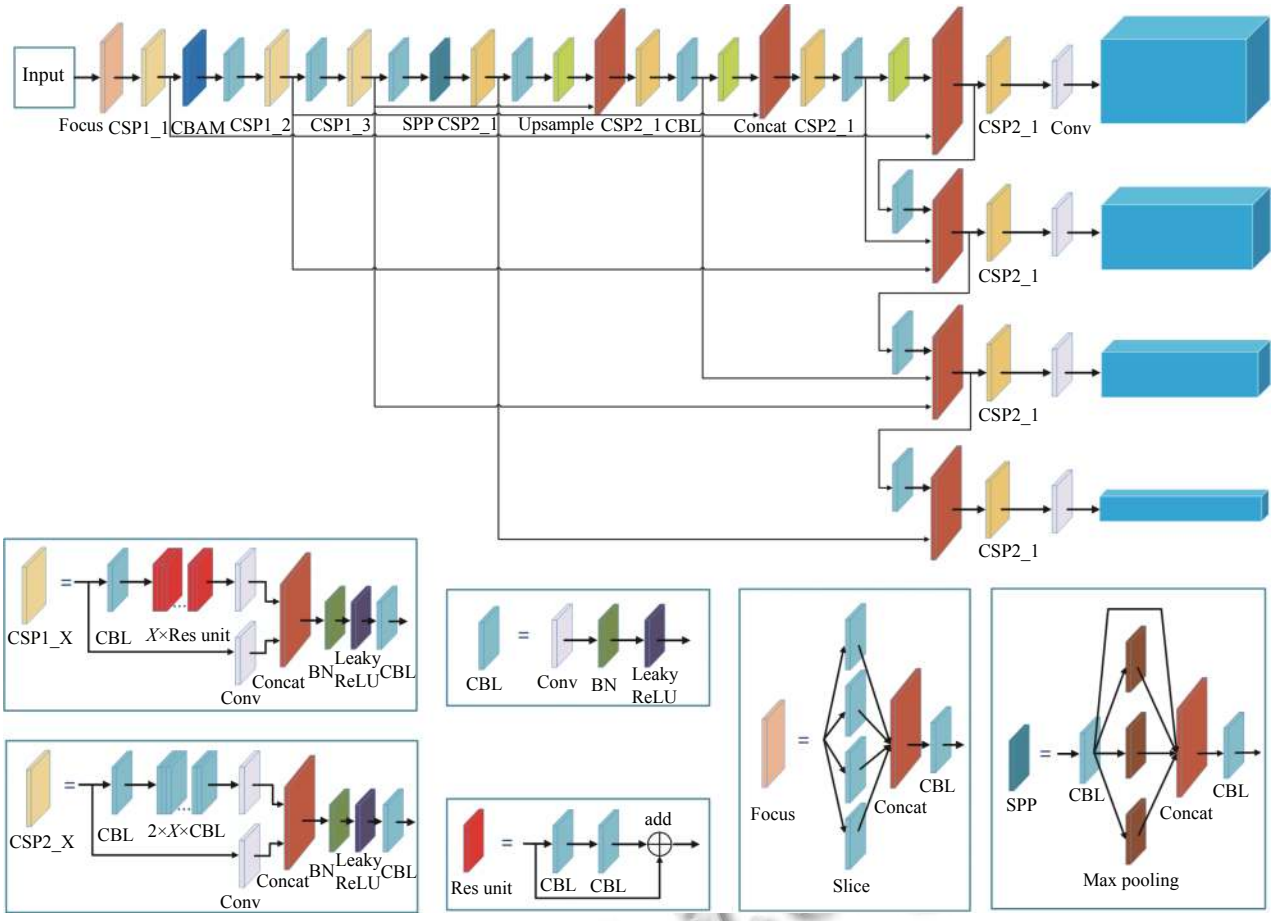


图 1 BCF-YOLOv5 模型结构图

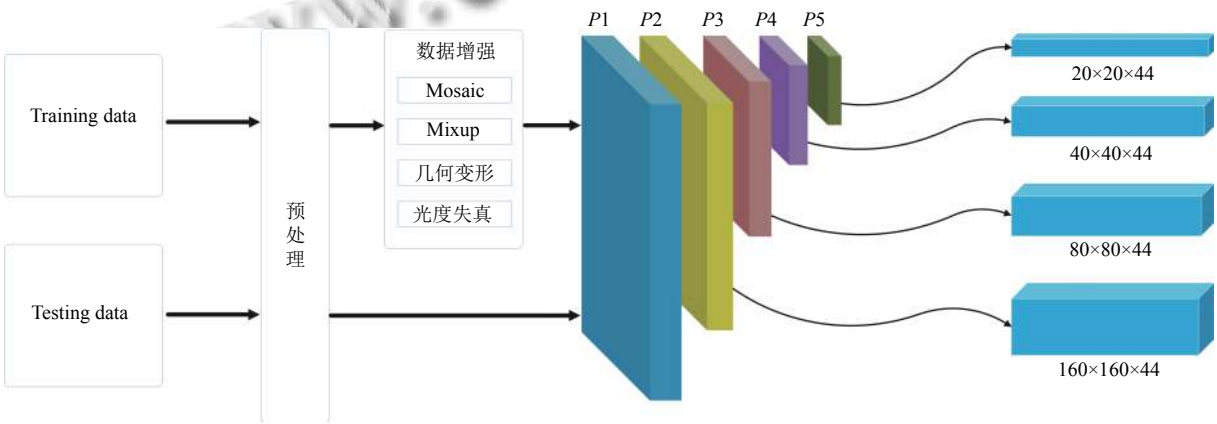


图 2 小目标检测头

1.2 CBAM 注意力模块

CBAM 是一种轻量的注意力模块, 其简单有效, 可以直接集成到 CNN 架构中, 并且可以端到端的对其进行训练. 在给定特征映射的情况下, CBAM 会依次沿着通道和空间两个独立维度推导注意映射, 然后将注意映射与输入特征映射相乘, 进行自适应特征细化. CBAM 模块的结构如图 3 所示, 在文献 [16] 中, CBAM 模块被集成到不同数据集和不同分类任务的不同模型中, 模型性能均得到了较大提升, 证明了 CBAM 模块的有效性.

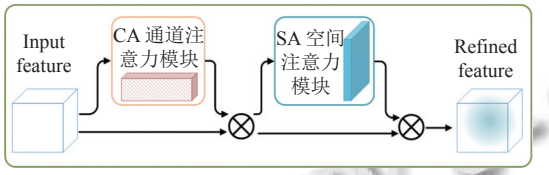


图 3 CBAM 注意力模块

小目标尺寸较小、特征少且不明显, 在主干网络中加入 CBAM 注意力模块, 可以增强网络对于目标特征的提取能力, 同时直接改善颈部网络部分的特征融合. 在检测任务中, CBAM 注意力模块可以帮助模型有效的提取注意区域, 提高检测性能.

1.3 BiFPN 网络

BiFPN[18] 网络是集成双向交叉连接和加权融合的一种高效的多尺度特征融合方法. 自 FPN[7] 被提出以来, FPN 被广泛应用于多尺度特征融合. 近些年来, PANet[19]、M2det[20] 和 NAS-FPN[21] 等更多的多尺度特征融合网络结构被研究学者们提出来, 但是在融合不同层次的输入特征时, 大部分工作都是不加区分的总结它们. 然而, 这些不同的输入特征有着不同分辨率, 对融合的输出特征具有在不同的贡献. 为此, 文献 [17] 提出了一种简单但高效的加权双向特征金字塔网络 (BiFPN), 它引入了可学习的权重因子来表征不同输入特征的重要程度, 同时反复应用自顶向下和自底向上的多尺度特征融合.

为充分利用目标的底层特征, 本文对颈部网络进行了改进, 将原始 PANet 网络换成 BiFPN 网络, 以提高检测精度, 其结构如图 4 所示. 尽管 YOLOv5 中的 PANet 通过自顶向下和自底向上的路径聚合实现了较好的多尺度特征融合结果, 但计算量较大, 且自底向上特征融合阶段的输入特征中并没有主干网络生成的原始输出特征. BiFPN 采用跨连接去除 PANet 中对特征融合贡献度较小的节点, 在同一尺度的输入节点和输出节点间增加一个跳跃连接, 在不增加较多成本的同时, 融合了更多的特征. 在同一特征尺度上, 把每一个双向路径看作一个特征网络层, 并多次反复利用同一层, 以实现更高层次的特征融合.

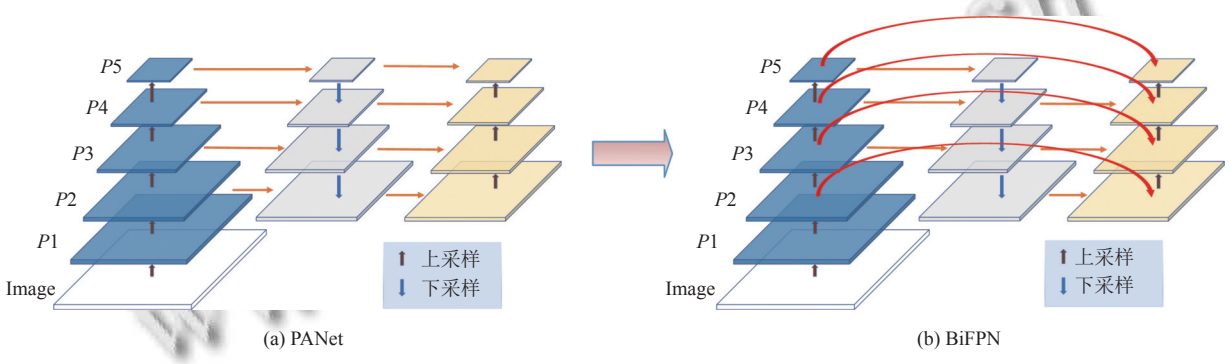


图 4 改进颈部网络

2 检测算法

本文通过加入 CBAM 注意力模块增强网络对于目标特征的提取能力, 搭配 BiFPN 结构和高分辨率检测头强化网络对于小目标的检测能力. 模型简化结构如图 5 所示, 其中$Pred4$是通过 $P2$ 层特征图H_2得到的高分辨率检测头, 其可以有效检测图像中的微小目标, 从而达到提升整体检测性能的目的.

2.1 加权特征融合

结合双向交叉连接和快速标准化特征融合. 为每个特征分支提供一个权重因子, 通过网络自学习得到最佳权值. 对于特征输入I, 快速标准化融合O的计算方式如式 (1):

$$O=\sum_i \frac{w_i}{\varepsilon+\sum_j w_j}\cdot I_i \tag{1}$$

其中, 通过 ReLU 函数保证每个w_i均满足$w_i \geqslant 0, \varepsilon = 0.0001$是一个极小值以避免数值不稳定.

本节以*Pred3*的计算过程为例, 以体现本文算法的计算过程. 计算过程如式 (2):

$$\begin{cases} H_2^{\text{out}'} = Upsample(H_2^{\text{out}}) \\ H_3^{\text{out}} = Conv\left(\dfrac{\omega_1 \cdot Conv^{1\times1}(I_3^{\text{out}}) + \omega_2 \cdot N_3^{\text{out}} + \omega_3 \cdot H_2^{\text{out}'}}{\omega_1 + \omega_2 + \omega_3 + \varepsilon}\right) \\ Pred3 = Conv(H_3^{\text{out}}) \end{cases} \tag{2}$$

其中, $Conv()$为卷积操作, $Upsample()$为线性 2 倍上采样操作, out表示该网络层输出. $Conv^{1\times1}(\cdot)$为1×1卷积, 调整I_3^{out}通道数使其与N_3^{out}和H_2^{out}相同, 以便可以进行特征融合.

2.2 Concat 特征融合

结合双向交叉连接和 Concat 特征融合. 为尽可能地保留特征, 不对通道维数进行压缩, 抛弃了加权融合的思想, 以一定的内存资源为代价保留完整的特征信息. 本节同样以*Pred3*计算为例, 过程如下:

$$\begin{cases} H_2^{\text{out}'} = Upsample(H_2^{\text{out}}) \\ H_3^{\text{out}} = Conv(Concat(I_3^{\text{out}} + N_3^{\text{out}} + H_2^{\text{out}'})) \\ Pred3 = Conv(H_3^{\text{out}}) \end{cases} \tag{3}$$

其中, $Concat(\cdot)$操作为将全部输入特征直接进行通道数叠加, 输出特征与各输入特征满足:

$$O_{\text{channel}} = \sum_i I_{\text{channel}}^i \tag{4}$$

其中, I表示输入特征, O表示输出特征, channel 表示通道数.

本文在 YOLOv5 中添加额外分支时提供了两种特征融合方案, 其中加权特征融合可以在同等资源下获取更为丰富的目标特征, 提高模型检测性能, 该特征融合策略可以用于轻量化模型中, 既可以保持较高的检测性能, 又可以节省较多的计算资源. 但是本文提出的 BCF-YOLOv5 模型支路过多 (支路数>2), 采用加权融合过度压缩了通道资源, 经实验得知其效果要低于 concat 特征融合方式, 尽管 concat 特征融合方式将占有更多的计算资源, 但本文为获得最优的检测性能, 依然采用 concat 特征融合方式.

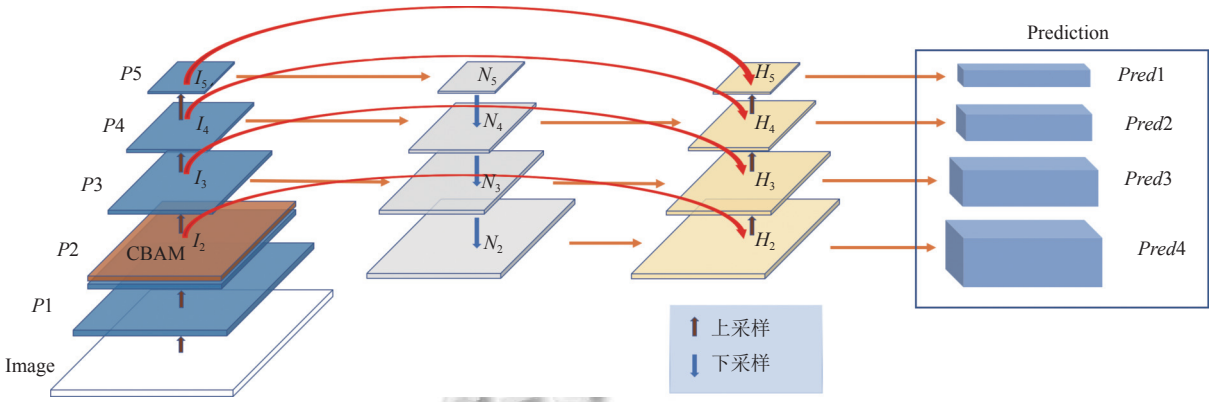


图 5 BCF-YOLOv5 简化结构图

3 实验

本文实验使用自建人脸瑕疵数据集与公开无人机数据集 Visdrone2019 评估模型性能, 并分别从数据集介绍、网络设置与训练、评价指标、实验结果分析和消融实验 5 个方面展开介绍.

3.1 数据集介绍

3.1.1 人脸瑕疵数据集

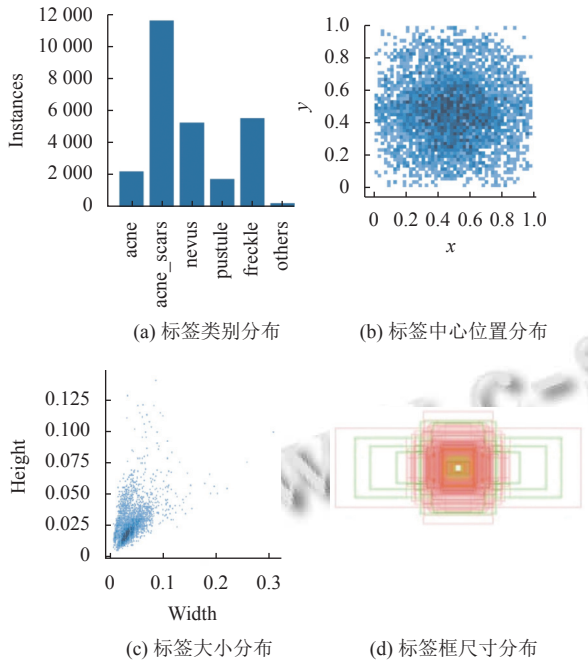
人脸瑕疵数据集由江苏医像信息技术有限公司提供, 全部使用专业设备 VISIA7 进行拍摄获取. 数据集包含高分辨率人脸图片 2 436 张, 标注的瑕疵类型共有 6 类, 分别为 acne (痤疮)、acne_scars (痘印)、nevus (黑痣)、pustule (脓疮肉囊)、freckle (色斑)、others (其他).

人脸瑕疵数据的标记通过 labelimg 手工标注完成, 瑕疵标签共有 33 218 个. 通过观察原始数据可以发现, 超过 45% 的瑕疵标签均在 16×16 像素以下. 数据可视化分析如图 6 所示.

3.1.2 VisDrone2019 公开无人机数据集

VisDrone2019 数据集由天津大学机器学习和数据挖掘实验室 AISKYEYE 团队收集, 基准数据集包含 288 个视频片段、261 908 帧和 10 209 幅静态图像. 数据集是在不同的场景、不同的天气和光照条件下使用各种无人机平台 (即不同型号的无人机) 收集的, 手工

标注了超过 260 万个边界框或经常感兴趣的目标点, 包括行人、人、轿车、货车、公共汽车、卡车、摩托车、自行车、遮阳篷三轮车和三轮车等 10 个感兴趣的对象类别. 其各类标签数量如表 1 所示.



(a) 标签类别分布 (b) 标签中心位置分布

(c) 标签大小分布 (d) 标签框尺寸分布

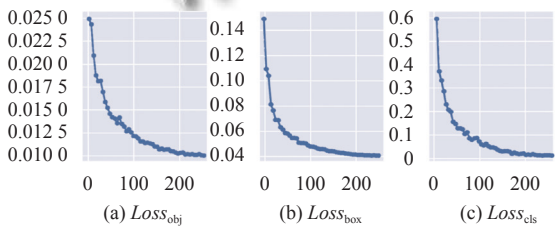
图 6 数据分析

3.2 网络设置与训练

本节将以人脸瑕疵数据集为例, 详细介绍 YOLOv5 与 BCF-YOLOv5 的训练过程. 实验所使用硬件配置为 24 GB 的 NVIDIA GeForce RTX 3090 显卡, 深度学习框架为 PyTorch 1.7.1, Python 版本为 3.8.5, CUDA 版本为 11.0, 操作系统为 Ubuntu 18.04.

表 1 VisDrone2019 数据集各类标签数量

种类	pedestrian	people	bicycle	car	van	trunk	tricycle	awning-tricycle	bus	motor
数量	79055	26962	10389	144620	24899	12875	4812	3245	5917	29618



(a) $Loss_{obj}$ (b) $Loss_{box}$ (c) $Loss_{cls}$

图 7 收敛图 (横坐标表示 epoch, 纵坐标表示损失函数)

由收敛图可知, 网络大约在 240 多 epoch 时, 即 73 000 余次迭代后各参数趋于平稳, 总损失函数下降到 0.06 左右. 从收敛情况来看, BCF-YOLOv5 网络的训练结果比较理想.

3.2.1 损失函数设置

BCF-YOLOv5 的损失函数与 YOLOv5 是一致的, 均包括置信度损失 ($Loss_{obj}$)、矩形框损失 ($Loss_{box}$) 和分类损失 ($Loss_{cls}$).

$$Loss = a \cdot Loss_{obj} + b \cdot Loss_{box} + c \cdot Loss_{cls} \quad (5)$$

其中, a、b、c均表示相应损失函数在总损失函数中的权重占比, 本文实验均取值 1, 表示三者权重一样. 其中分类和置信度损失均使用 BCEWithLogitsLoss 函数, 矩形框回归采用 CIOU_Loss.

3.2.2 网络训练

在训练前, 将数据集图片和标签按照 8:1:1 的比例分为训练集、验证集和测试集. 在训练阶段, 我们首先训练原始 YOLOv5 模型, 并使用了官方预训练模型, 这样可以加速网络训练. 在训练数据集上最大训练次数 epoch 为 250, 前 3 个 epoch 用于热身训练, 并采用 SDG 优化策略进行学习率调整, 使用 3E–2 作为余弦退火策略的初始学习率, 最后一个 epoch 学习率降为 3E–3, 即初始学习率的 0.1, 训练过程中 batch_size 设置为 8. YOLOv5 模型训练完成以后, 开始训练 BCF-YOLOv5 模型, 这里将使用前面 YOLOv5 在人脸瑕疵数据集上的部分训练权重, 因为 BCF-YOLOv5 与 YOLOv5 共享骨干网络的大部分, 使用这些权重, 我们可以将许多权重从 YOLOv5 转移到我们新的模型上, 可以节省大部分的训练时间. 同样的, 训练最大次数 epoch 依然是 250, batch_ size 设置为 8, 超参数全部保持一致.

BCF-YOLOv5 网络训练过程中各收敛曲线如图 7 所示, 对各损失函数分别进行分析, 可以确保总损失函数的收敛完整度.

3.3 评价指标

为验证模型性能, 本文选用精确率 (Precision, P)、召回率 (Recall, R)、平均精度 (average precision, AP)、平均精度均值 (mean average precision, mAP) 来评估模型的检测性能.

1) 精确率和召回率:

$$P = \frac{TP}{TP + FP} \cdot 100\% \quad (6)$$

$$R = \frac{TP}{TP+FN} \cdot 100\% \tag{7}$$

其中, *TP* (true positives) 表示被正确检测出来的目标数量, *FP* (false positives) 表示被检测为目标的背景数量, *FN* (false negatives) 表示被检测为背景的目标数量.

2) 平均精度和平均精度均值:

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{\sum P_A}{N_C} \tag{9}$$

其中, N_C为类别数量, P_A为各类别的平均精度. 利用实验数据可绘制模型的 PR 曲线, 曲线所围面积即为 *AP*, 该指标被用来评估模型对于单个类别的目标检测性能表现, 将所有类别的 *AP* 值取平均就得到了 *mAP*. *mAP* 的值在 0–1 之间, *mAP* 值越接近于 1 表示模型的性能越好, 检测能力越强.

3.4 实验结果与分析

3.4.1 人脸瑕疵数据集实验分析

为验证本文算法的有效性, 基于人脸瑕疵数据集, 选择 YOLOv3、RetinaNet[22]、CornerNet[23] 和 YOLOv5 作为对比网络模型. 实验结果如表 2 所示. 由表 2 可知, BCF-YOLOv5 在单类 *AP* 上和 *mAP* 两个指标上, 都远高于 YOLOv3、RetinaNet 和 CornerNet. 与原始 YOLOv5 对比, 各类 *AP* 均有一定的提升, 在 *mAP* 上提高 7.26%, 证明了 BCF-YOLOv5 模型对于小目标具有更强的检测能力. YOLOv5 和 BCF-YOLOv5在人脸瑕疵数据集上的 *mAP* 曲线如图 8 所示. 在图 9 中本文展示了一些人脸瑕疵图片的检测结果图片, 可以看出本文算法对人脸瑕疵可以达到较好的检测效果.

表 2 人脸瑕疵数据集实验结果 (%)

模型	*AP*						*mAP*
	acne	acne_scars	nevus	pustule	freckle	others	
YOLOv3	30.7	33.5	42.2	62.3	23.1	21.5	35.55
RetinaNet	34.8	41.3	49.9	65.4	21.3	45.7	44.07
CornerNet	38.1	45.8	48.2	67.1	25.6	48.3	45.52
YOLOv5	56.0	57.6	72.4	79.3	41.9	55.9	60.52
BCF-YOLOv5 (本文)	58.3	60.6	77.9	90.3	52.6	67.0	**67.78**

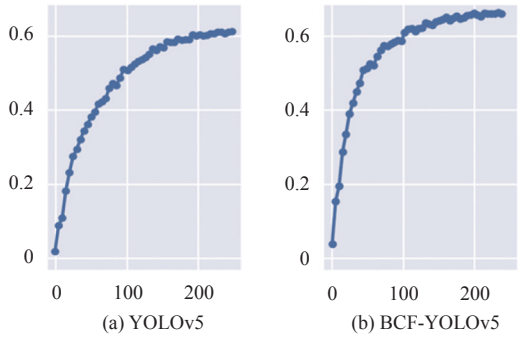


图 8 *mAP* 曲线 (横坐标表示 epoch, 纵坐标表示 *mAP*@0.5)

图 9 BCF-YOLOv5 在人脸瑕疵测试集上的一些检测结果

3.4.2 VisDrone2019 无人机数据集实验分析

为保证实验的严谨性与可对比性, 本文实验参考文献 [24] 的实验设置, 在 VisDrone2019 上分别验证了 YOLOv3、YOLOv5 和 BCF-YOLOv5, 同时取文献 [24] 中 BetterFPN 和 RRNet 的实验结果进行对比, 对比结果如表 3 所示.

由表 3 可知, BCF-YOLOv5 在 *mAP* 指标上远高于 YOLOv3、BetterFPN 和 RRNet, 较 YOLOv5 提高 1.63%, 实验结果表明了 BCF-YOLOv5 模型在小目标数据集上的有效性以及在不同数据集上也具有较强的

泛化能力. BCF-YOLOv5 在 VisDrone2019 数据集上训练的混淆矩阵如图 10 所示. 图 11 展示了 VisDrone2019 数据集的一些检测可视化结果.

表 3 VisDrone2019 无人机数据集实验结果

模型	*mAP* (%)
YOLOv3	20.41
RRNet[24]	29.13
BetterFPN[24]	28.55
YOLOv5	39.73
BCF-YOLOv5 (本文)	**41.36**

3.5 消融实验

本节基于人脸瑕疵数据集, 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. 以原始 YOLOv5 为基础, 依次增加各模块进行实验, 表 4 列出了实验结果.

由表 4 可知, 增加 *P*2 层检测头对于模型的提升较大, *mAP* 提升 1.71%, 说明增加高分辨率检测头可以有效增强对于微小目标的检测能力. 同时可以发现, 单独增加 CBAM 和 BiFPN 模块对于模型的提升较为一般, *mAP* 分别提升 0.98% 和 1.66%, 同时增加两个模块时, *mAP* 提升 5.55%, 说明主干网络特征提取的质量对颈部网络的特征融合产生了较大影响, CBAM 注意力模块有效提升了模型的特征提取能力, 结合 BiFPN 网络的特征融合, 有效提升了模型的检测性能.

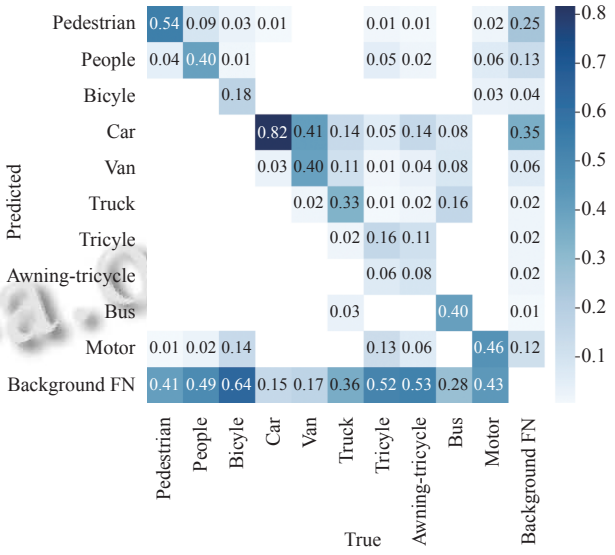


图 10 VisDrone2019 数据训练混淆矩阵

图 11 BCF-YOLOv5 在 VisDrone2019 数据测试集上的一些检测结果

表 4 BCF-YOLOv5 消融实验结果

模型	*mAP* (%)
YOLOv5x	60.52
YOLOv5x+P2	62.23
YOLOv5x+P2+CBAM	63.21
YOLOv5x+P2+BiFPN	63.89
YOLOv5x+P2+CBAM+BiFPN	**67.78**

4 结论与展望

针对小目标难以检测的问题, 本文提出了一种改进的 YOLOv5 检测模型, 在主干网络中添加 CBAM 注意力模块, 提高网络的特征提取能力; 在颈部网络部分将 PANet 网络结构替换成 BiFPN, 强化底层特征利用; 最后在检测头部分增加一个高分辨率检测头用于检测尺寸过小的目标, 有效提高了对模型对于小目标的检测精确度. 本文在人脸瑕疵数据集和 VisDrone2019 无人机数据集上做了多组对比实验, 在 mAP 上均得到了一定提升, 证明了本文方法的有效性. 不过, 由于模型中增加了额外的检测头, 同时使用 concat 特征融合, 增加了模型的大小, 后续工作就致力于提升检测精度的同时减小模型大小.

参考文献

1 Wu XW, Sahoo D, Hoi SCH. Recent advances in deep learning for object detection. Neurocomputing, 2020, 396: 39–64. [doi: 10.1016/j.neucom.2020.01.085]

2 Singh B, Najibi M, Davis LS. SNIPER: Efficient multi-scale training. Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montréal: Curran Associates Inc., 2018. 9333–9343.

3 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2012. 1097–1105.

4 Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587.

5 Girshick R. Fast R-CNN. 2015 IEEE International Conference on Computer Vision (ICCV). Santiago: IEEE, 2015. 1440–1448.

6 Ren SQ, He KM, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.

7 Lin TY, Dollár P, Girshick R, *et al*. Feature pyramid networks for object detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 936–944.

8 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot MultiBox detector. Proceedings of the 14th European Conference on Computer Vision. Amsterdam: Springer, 2016. 21–37.

9 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

10 Redmon J, Farhadi A. YOLO9000: Better, faster, stronger. 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017. 6517–6525.

11 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

12 Song XY, Gu W. Multi-objective real-time vehicle detection method based on YOLOv5. 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM). Xi'an: IEEE, 2021. 142–145. [doi: 10.1109/ISAIAM53259.2021.00037]

13 Sruthi MS, Poovathingal MJ, Nandana VN, *et al*. YOLOv5 based Open-Source UAV for Human Detection during Search and Rescue (SAR). 10th International Conference on Advances in Computing and Communications. Kochi: IEEE. 2021. 1–6.

14 Dong X, Xu NN, Zhang LY, *et al*. An improved YOLOv5 network for lung nodule detection. 2021 International Conference on Electronic Information Engineering and Computer Science (EIECS). Changchun: IEEE, 2021. 733–736. [doi: 10.1109/EIECS53707.2021.9588065]

15 Zhu XK, Lyu SC, Wang X, *et al*. TPH-YOLOv5: Improved YOLOv5 based on Transformer prediction head for object detection on drone-captured scenarios. Proceedings of the IEEE International Conference on Computer Vision. Online: IEEE. 2021. 2778–2788.

16 Woo S, Park J, Lee JY, *et al*. CBAM: Convolutional block attention module. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 3–19. [doi: 10.1007/978-3-030-01234-2_1]

17 Bai HY, Wen S, Chan SHG. Crowd counting on images with scale variation and isolated clusters. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul: IEEE, 2019. 18–27. [doi: 10.1109/ICCVW.2019.00009]

18 Tan MX, Pang RM, Le QV. EfficientDet: Scalable and efficient object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle: IEEE, 2020. 10778–10787. [doi: 10.1109/CVPR42600.2020.01079]

19 Liu S, Qi L, Qin HF, *et al*. Path aggregation network for instance segmentation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018. 8759–8768. [doi: 10.1109/CVPR.2018.00913]

20 Zhao QJ, Sheng T, Wang YT, *et al*. M2Det: A single-shot object detector based on multi-level feature pyramid network. Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu: AAAI, 2019. 9259–9266.

21 Ghiasi G, Lin TY, Le QV. NAS-FPN: Learning scalable feature pyramid architecture for object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019. 7029–7038.

22 Lin TY, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2999–3007.

23 Law H, Deng J. CornerNet: Detecting objects as paired keypoints. Proceedings of the 15th European Conference on Computer Vision. Munich: Springer, 2018. 765–781.

24 Zhu PF, Wen LY, Du DW, *et al*. Detection and tracking meet drones challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021: 1–1. [doi: 10.1109/TPAMI.2021.3119563]

(校对责编: 牛欣悦)