Computer Science and Application 计算机科学与应用, 2022, 12(6), 1700-1709
Published Online June 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.126170

Hans汉斯

基于YOLOv3的轻量化口罩检测算法研究

张寿明[1]，刘　凯[1,2]

[1]昆明理工大学信息工程与自动化学院，云南　昆明
[2]昆明理工大学云南省人工智能重点实验室，云南　昆明

收稿日期：2022年5月26日；录用日期：2022年6月23日；发布日期：2022年6月30日

摘　要

针对当前基于深度学习的口罩检测算法在实时性与检测精度上不能同时具有良好的表现性能，本文提出一种基于YOLOv3的轻量化口罩检测算法，通过EfficientNet-B1网络替换掉原有的网络参数量大，网络结构复杂的骨干网络Darknet-53，为进一步提升网络性能，实验引入ECA通道注意力机制与特征金字塔结构相结合，最后采用CIOU对原有的边界框损失进行优化。实验结果表明，该网络结构模型与YOLOv3相比，检测精度仅降低1.73%，但模型参数量降低了79%，且单张图片检测速度也提升了3.93倍，一定程度上体现了本文算法的良好性能。

关键词

深度学习，目标检测，轻量化，YOLOv3，通道注意力机制，CIOU

Research on Lightweight Mask Detection Algorithm Based on YOLOv3

Shouming Zhang[1], Kai Liu[1,2]

[1]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan
[2]Key Laboratory of Artificial Intelligence in Yunnan Province, Kunming University of Science and Technology, Kunming Yunnan

Received: May 26th, 2022; accepted: Jun. 23rd, 2022; published: Jun. 30th, 2022

Abstract

In view of the fact that the current deep learning-based mask detection algorithm cannot have good performance in real-time and detection accuracy at the same time, this thesis proposes a

文章引用: 张寿明, 刘凯. 基于 YOLOv3 的轻量化口罩检测算法研究[J]. 计算机科学与应用, 2022, 12(6): 1700-1709.
DOI: 10.12677/csa.2022.126170

lightweight mask detection algorithm based on YOLOv3, which replaces the original backbone network Darknet-53 which has a large number of network parameters and a complex network structure through the EfficientNet-B1 network. In order to further improve the network performance, the experiment introduces the ECA channel attention mechanism combined with the feature pyramid structure, and finally uses CIOU to optimize the original bounding box loss. The experimental results show that the network structure model is compared with YOLOv3. The detection accuracy is only reduced by 1.73%, but the amount of model parameters is reduced by 79%, and the detection speed of a single image is also increased by 3.93 times, which reflects the good performance of the algorithm in this paper to a certain extent.

Keywords

Deep Learning, Object Detection, Lightweight, YOLOv3, Channel Attention Mechanism, CIOU

1. 引言

自 2019 年 12 月以来，新型冠状病毒席卷全球[1]，对人民群众的日常生活乃至生命财产安全都造成了重大影响。尽管当前的防疫措施已逐渐趋于完善，但对人力物力等资源的消耗也是显而易见的。随着计算机视觉和嵌入式系统的迅速发展，两者应用开始结合，并极大方便了人们的生活，例如：人脸抓拍相机、人脸识别考勤机、以及停车场进出口车牌识别系统等等。因此将目标检测网络部署在对应的硬件平台上，并将系统应用在人流量较大的公众场所，就能很好地替代部分人工进行复杂场景下的人脸口罩检测，这种方案能一定程度上提高了人们的出行效率，减少人群聚集，满足国家提出的防疫要求。基于目标检测算法的口罩检测系统如图 1 所示：

Figure 1. Mask detection system
图 1. 口罩检测系统

该系统工作流程是先由图像采集设备将实时采集到的图片传入已经部署相应算法的嵌入式平台中，然后将处理后的结果传入显示器，由安排的工作人员进行监管。由于平台算力和成本等综合因素，这一系统落地的关键就在于算法性能的优化。

目标检测[2] [3]的发展脉络可以将其划分成为两个周期：传统目标检测算法时期(1988 年~2014 年)和基于深度学习的目标检测算法时期，且基于深度学习的目标检测算法可以分为两大类，一是以 SPPNet、RCNN 系列为代表的 two-stage 检测方法，二则是以 SSD 和 YOLO 系列为代表的 one-stage 检测方法。前者的检测流程主要分为两个阶段：先从图像中通过候选区域最后生成物体的边框。而后者的检测不需要候选区域就能直接得到物体的类别概率和位置坐标值，因此在实时性上相对前者有着较为明显的优势。

经过此次疫情影响，许多学者展开了对人脸佩戴口罩的相关算法展开了研究，文献[4]在 YOLOv3 [5]算法的基础上对特征融合部分进行了改进，通过添加浅层特征图形成 4 尺度检测结构，并引入自上而下和自下而上的多尺度融合结构，进一步利用了特征信息，有效提升了检测精度，但一定程度上也增加了模型参数量。文献[6]以 SSD [7]网络为基础，引入特征融合网络有效增强算法对细节信息学习和处理能力，并利用 Quality Focal Loss 损失函数调节正负样本的权值，实验表明改进后的算法精度得到了小幅提高，但对表征信息较弱的目标检测效果不佳。文献[8]设计出基于 YOLOv3 网络引入 SPPNet 结构的口罩检测算法，以空间金字塔的网络结构更好地融合特征信息，虽然增大了网络参数导致检测速度受到了一定影响，但网络特征提取与融合能力得到明显提升，同时将跨阶段局部网络引入特征提取部分，一定程度上降低了计算量，且测试的效果 Map0.5 达到了 90%，但损失函数部分未进行改进，实际检测中边界框存在问题。

鉴于当前针对复杂场景下的人脸口罩佩戴检测算法在性能上各有千秋，但为了更好地将生成的网络模型部署在硬件设备上投入于公共服务上，将现有的目标检测网络进行轻量化改进更是当前所需的。本文主要研究内容如下：本研究提出基于 YOLOv3 的轻量化口罩检测算法。

2. YOLOv3 网络

本文主要是对 YOLOv3 网络进行轻量化改进。其网络结构如图 2 所示。

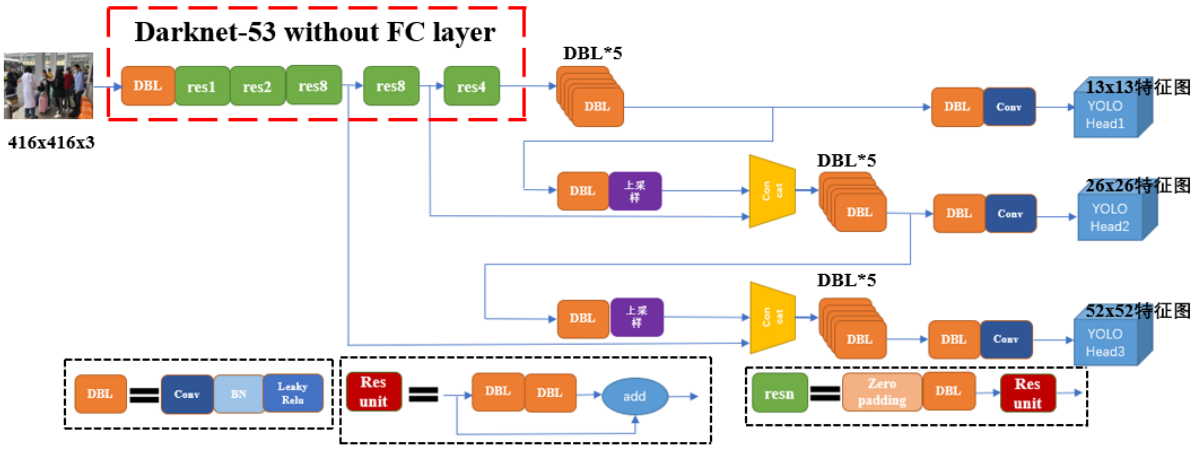

Figure 2. Network structure diagram of YOLOv3
图 2. YOLOv3 网络结构图

YOLOv3 使用 Darknet-53 作为主干特征提取网络提取三个不同尺度的有效特征层，通过构建特征金字塔结构实现不同尺度之间的特征融合并进行特征提取，最后利用多尺度检测进行目标预测。

Darknet-53 主要由 53 个基本单元 DBL [9]构成，其中每个 DBL 结构都包含了 1 × 1 和 3 × 3 的卷积层，批量归一化层，以及 Leaky ReLU [10]激活函数层。与上一代 YOLOv2 的骨干网络 Darknet-19 对比，Darknet-53 主要做了如下改进：1) 没有采用最大池化层，转而采用步长为 2 的卷积层进行 5 次下采样，以此减少因池化造成特征信息的丢失。2) 为了防止过拟合，在每个卷积层之后加入了一个 BN 层和一个 Leaky ReLU 激活函数。3) 引入了残差网络。并在网络结构中使用大量的残差块，有效将网络更深层的特征提取出来，同时避免训练中产生梯度消失以及梯度爆炸等问题。

3. 改进的 YOLOv3 网络

3.1. 轻量化改进

轻量化网络成为近年来许多学者的一个研究热点，轻量化网络顾名思义就是网络的参数量比较少，计算量也相对较小，目前常用的一些减少网络计算量的方法分为以下几种：1) 基于轻量化网络设计：比如 EfficientNet 系列，Mobilenet 系列，Shufflenet 系列，Xception [11]等，使用分组卷积[12] (Group Convolutional)、1 × 1 卷积[13]在一定程度上保证网络精度模型模型同时减少了计算量。2) 模型剪枝[14]：常常用在参数量足够大的网络，在大型网络训练中通常存在计算量的冗余，模型剪枝的原理就是减少网络的冗余部分，进而减少网络的计算量。3) 量化：利用 TensorRT 量化，一般在 GPU 上可以提速几倍。4) 知识蒸馏[15]：利用大模型来帮助小模型学习，提高小模型的精度。

本文提出一种基于 YOLOv3 改进的轻量级目标检测算法，EfficientNet [16]网络是 2019 年由 Google 团队提出，以 MnasNet 的基本模块 MBConv 为搜索空间，搜索出了一个基准网络 EfficientNet-B0，并将其作为基准网络探究出 EfficientNetB1~B7 共 7 种网络结构。其中 EfficientNet-B1 网络结构如表 1 所示。

Table 1. The network structure of EfficientNet-B1
表 1. EfficientNet-B1 网络结构

结构	操作	分辨率	输出通道数	堆叠数	步距
1	Conv3 × 3	224 × 224	32	1	2
2	MBconv1, k3 × 3	112 × 112	16	2	1
3	MBconv6, k3 × 3	112 × 112	24	3	2
4	MBconv6, k5 × 5	56 × 56	40	3	2
5	MBconv6, k3 × 3	28 × 28	80	4	2
6	MBconv6, k5 × 5	14 × 14	112	4	1
7	MBconv6, k5 × 5	14 × 14	192	5	2
8	MBconv6, k3 × 3	7 × 7	320	2	1
9	Conv1 × 1&Pooling&FC	7 × 7	1280	1	——

EfficientNet-B1 网络结构主要是由 MBconv 结构堆叠而成，其网络结构如图 3 所示，

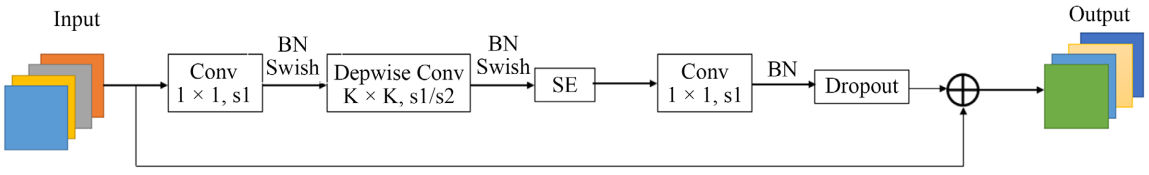

Figure 3. Network structure diagram of MBConv
图 3. MBConv 网络结构图

由上图可以看出，该网络的输入经过主分支的 1 × 1 的卷积进行升维，后经过 BN 层加 swish 激活函数后进行逐通道卷积，所得到的特征经过 SE 通道注意力机制后通过 1 × 1 的卷积进行降维处理后经 dropout 比率为 0.2 的正则化，当输入与输出的 shape 相同时，则会出现对应的 shortcut 连接。

原 YOLOv3 网络在进行骨干网络特征提取时，需要执行 5 组残差块，且每组均含有一次步长为 2 的卷积操作，即 Darknet-53 进行了连续 5 次的下采样，其中将后 3 次下采样结果作为骨干网络的输出部分，传入到网络的特征融合部分中进行后续的特征处理。而 EfficientNet-B1 将输入图片传入进网络中也进行了 5 次步距为 2 的下采样处理，使得特征图的宽高得到了 5 次压缩。同理本实验将后 3 次特征图的压缩结果替代为原 YOLOv3 骨干网络的输出部分传入到网络的特征融合部分。

3.2. 特征金字塔结构的改进

原有的网络特征融合部分是将从骨干网络输入进来的三个不同尺度的特征图通过自下而上的特征金字塔结构进行网络特征融合，将深层的网络信息利用上采样和堆叠对整个金字塔结构语义信息进行增强，后续进行多尺度预测以得到良好的检测效果。原有的特征金字塔结构示意图如图 4 所示。

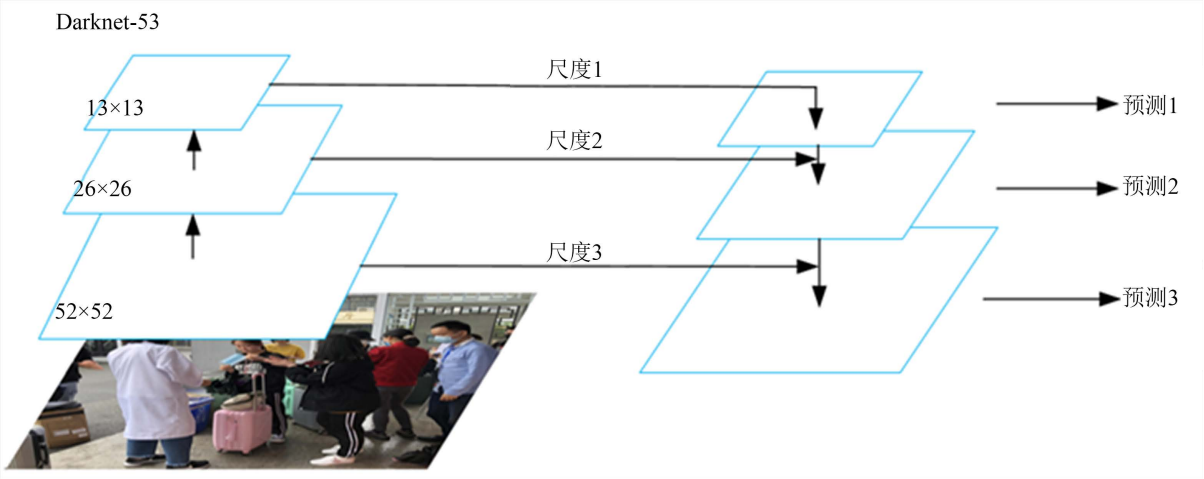

Figure 4. The original feature pyramid structure diagram
图 4. 原有的特征金字塔结构图

为进一步提升网络性能，实验将注意力机制与特征金字塔结构融合，使网络聚焦于图片的有效区域，而不是关注图片内的所有像素点。注意力机制作为机器学习常见的数据处理方法，能有效的提升网络的聚焦能力，其中 SENet [17]，ECANet [18]为常用的通道注意力机制，其网络结构分别如图 5，图 6 所示。

在 SENet 中输入特征首先会逐通道经过全局均值池化，随后经过两层全连接层，最后经过 Sigmoid 非线性激活后产生每一通道的权重。这两层 FC 的作用就是捕获跨通道之间的非线性交互，可以有效降低维度，但降维不可避免会带来副作用，而且不利于捕获通道之间的依赖关系。ECANet 在经过 SENet 的全局均值池化后，考虑到每个通道及其 k 个近邻，通过一维卷积快速完成通道权重的计算。k 代表在一个通道权重的计算过程中参与的近邻数目，其大小影响 ECA 计算的效率和有效性。为此实验使用自适应函数用来计算 k 的值。

$$k = \varphi_c = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd} \tag{1}$$

其中 C 为通道数，γ 和 b 分别取 2 和 1。

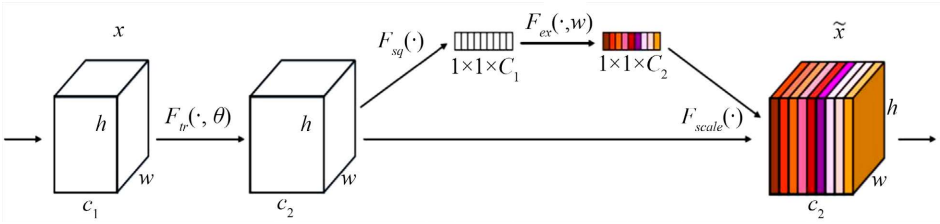

Figure 5. Network structure diagram of SENet
图 5. SENet 网络结构图

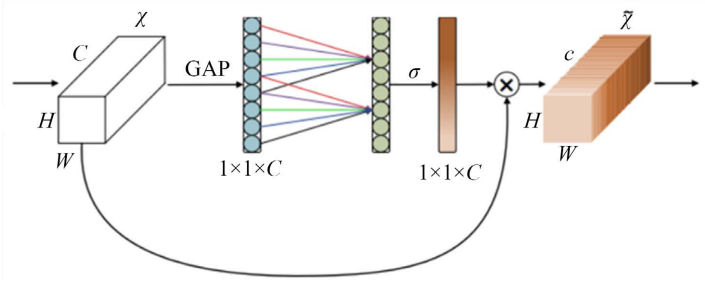

Figure 6. Network structure diagram of ECANet
图 6. ECANet 网络结构图

3.3. 边界框损失优化

交并比(IOU)是目标检测中最常用的指标，其作用不仅用来确定正样本和负样本，还可以用来评价预测框和真实框的检测效果。其公式为：

$$\mathrm{IOU} = S_{交} / S_{并} \tag{2}$$

IOU 是体现预测检测框与真实检测框的重合程度，但在网络训练中可能会出现预测框和真实框无重叠部分的情况，导致计算得到的 IOU 的值为 0，影响网络训练中参数后续的学习，而且 IOU 也无法精确地反映两者的重合度大小。如下图 7 所示，三种场景下计算出的 IOU 都相等，但实际预测框的回归效果存在差异，且左边的预测框回归的效果最好，右边的最差。

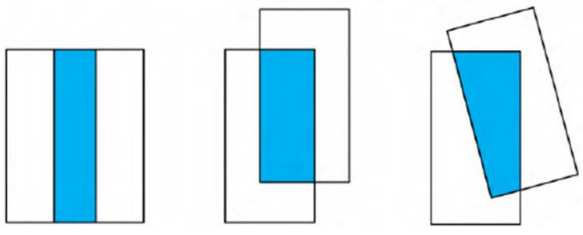

Figure 7. Comparison diagram of the same IOU
图 7. 相同 IOU 的对比示意图

CIOU 作为一个优秀的回归定位损失不仅考虑了两框的重合面积，还将中心点坐标距离和长宽比也加入了定位回归的考虑因素中，使网络在训练时的边界框损失收敛速度和最终生成的模型性能都能得到一定的提高。其公式为：

$$\mathrm{CIOU} = \mathrm{IOU} - \left(\frac{\rho^2\left(b^p - b^g\right)}{c^2} + \alpha v \right) \tag{3}$$

上式中 b^p、b^g 分别代表了预测框和真实框的中心点，ρ 代表的是计算两个中心点间的欧式距离，c 代表的是能够同时包含预测框和真实框的最小闭包区域的对角线距离，α 为权重函数，v 是用来度量宽高比的一致性，两者对应的公式如下：

$$\alpha = \frac{v}{1-\mathrm{IOU}+v} \tag{4}$$

$$v = \frac{4}{\pi^2}\left(\arctan w_g h_g - \arctan w_p h_p\right) \tag{5}$$

上式的 w_g，h_g，w_p，h_p 分别为真实框的宽高和预测框的宽高。CIOU 对应的损失函数公式为：

$$\mathrm{Loss}_{\mathrm{CIOU}} = 1-\mathrm{IOU}+\frac{\rho^2\left(b^p-b^g\right)}{c^2}+\alpha v \tag{6}$$

3.4. 改进的 YOLOv3 网络结构

本文提出的改进后的 YOLOv3 算法如图 8 所示，以 EfficientNet-B1 替代原 YOLOv3 的骨干网络 Darknet-53，将后三次下采样的结果 40 × 52 × 52，112 × 26 × 26，320 × 13 × 13 传入后续的特征网络结构中，通过 ECANet 和特征金字塔融合结构进一步提高网络模型性能，并采用 CIOU 作为边界框损失进行模型训练，完成网络反向传播和参数迭代更新。

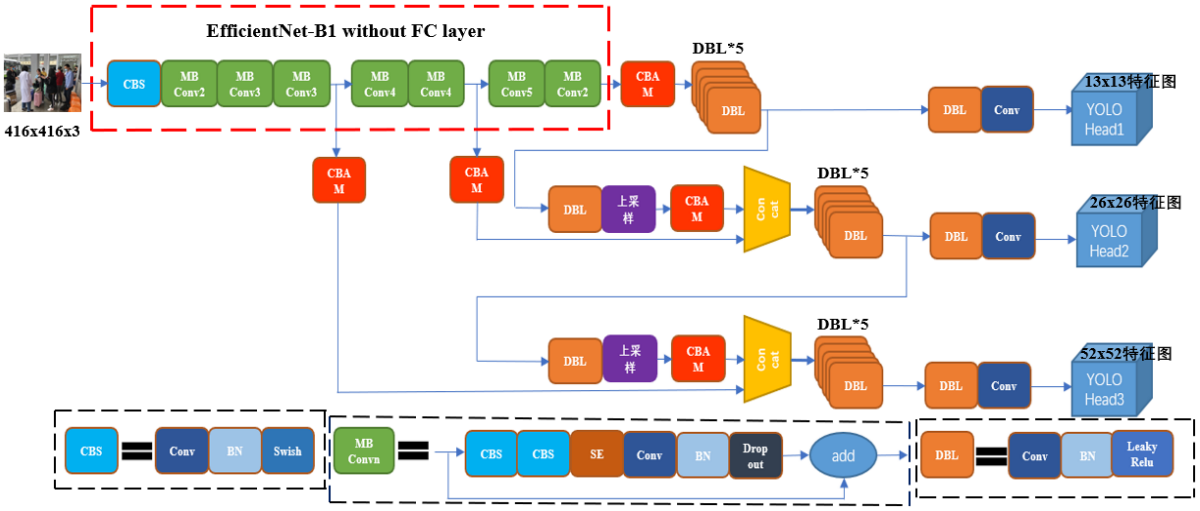

Figure 8. The improved YOLOv3 network structure diagram
图 8. 改进的 YOLOv3 网络结构图

4. 实验结果分析

4.1. 数据集建立和环境搭建

本文的数据集通过网络爬取的方式获得相应的数据集后经过清洗，筛选后共获得复杂场景下共 7607 张口罩数据集。采用 Labelimg 进行数据标注，在 Ubuntu18.04 的操作系统中进行训练，GPU 为 Nvidia RTX 3060，显存为 12 G，CUDA 版本为 11.4，深度学习框架为 pytorch 1.90。

4.2. 网络评价指标

本实验采取平均精度(Average Precision, AP)，平均精度均值(Mean Average Precision, MAP)以及单张

图片检测速度(Time)作为网络模型的评价指标。其中 AP 和 MAP 对应的公式为：

$$\mathrm{AP}=\int_0^1 P(R)\mathrm{d}R \tag{7}$$

$$\mathrm{Map}=\frac{\sum_{i=1}^{N}\mathrm{AP}_i}{N} \tag{8}$$

其中，N 表示检测目标的类别数，平均精度可以通过 P-R 曲线，体现对应类别的查准率(Precision)和查全率(Recall)。其公式如下：

$$\mathrm{Precision}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{9}$$

$$\mathrm{Recall}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{10}$$

上式中 TP 表示正样本被模型预测为正样本的数量，FN 表示正样本被预测为负样本的数量，FP 表示负样本被模型预测为正样本的数量。

4.3. 实验结果及分析

实验在网络训练中以 1:9 的比例划分训练集和验证集，采用 Adam 优化器进行网络优化，共迭代 500 轮，模型训练时长为 9 小时，最终生成的 AP 曲线如下图 9 所示。

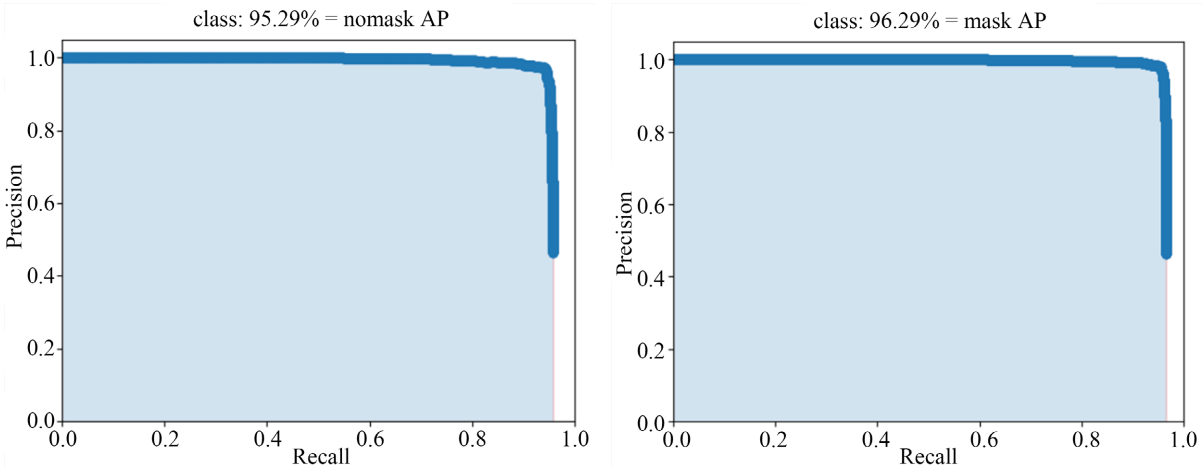

Figure 9. AP graph of the algorithm
图 9. 算法 AP 曲线图

为验证本文算法的有效性，在现有数据集上采用消融实验得到的实验结果如下表 2 所示。其中，方法 1 为 EfficientNet-B1 网络替换原 YOLOv3 骨干网络 Darknet-53 后的实验结果，方法 2 是基于方法 1 后的特征金字塔结构改进后的实验结果，方法 3 即为本文算法，即在方法 2 的基础上对边界框损失进行优化。

由表 2 可以看出本文算法依次改进的有效性，本文算法模型对比原 YOLOv3 网络模型性能，在牺牲 1.71%的检测精度下，参数量降为原 YOLOv3 的 21%，且检测速度提升 3.93 倍，很好地平衡了模型的检测精度和实时性。图 10 为本文算法的实际部分测试图，可以看出本文算法在不同的复杂环境下均能有效检测出口罩的佩戴情况，同时体现了本文改进的可行性。

Table 2. Ablation experiment results
表 2. 消融实验结果

	AP (%)		MAP (%)	#Params (MB)	Time (s)
	Nomask	Mask			
YOLOv3	96.99	98.11	97.50	235.06	0.1650
方法 1	93.58	94.46	94.02	50.30	0.0417
方法 2	94.53	95.43	94.98	50.35	0.0420
方法 3 (本文算法)	95.29	96.29	95.79	50.35	0.0420

Figure 10. Mask detection renderings
图 10. 口罩检测效果图

为更好地验证本文算法的有效性，在相同的数据集下，与当前主流的 SSD，YOLOv4 目标检测算法进行检测精度和实时性的对比，实验结果如表 3 所示。

Table 3. Performance comparison of different algorithms
表 3. 不同算法的性能对比

	AP (%)		MAP (%)	Time (s)
	Nomask	Mask		
YOLOv3	96.99	98.11	97.50	0.1650
SSD	92.54	93.82	93.18	0.1329
YOLOv4	97.36	98.72	98.04	0.1967
本文算法	95.29	96.29	95.79	0.0420

由表 3 可知，与当前主流目标检测算法相比，本文算法在性能的平衡上仍有着一定的优势，能更好

地应用于复杂场景下的实时人脸口罩的检测。

5. 结论

本文提出了基于 YOLOv3 的轻量级口罩检测算法，以 Efficient-B1 替代原 YOLOv3 网络结构复杂的 Darknet-53 骨干网络，并将特征金字塔结构与 ECANet 结合，通过自上而下的特征信息融合以及通道注意力机制使网络训练时更加关注样本的有效区域，进一步提升模型性能，最后采用 CIOU 对原有的边界框损失进行优化。实验结果表明，本文算法与原 YOLOv3 相比，在降低 1.73%检测精度的条件下，模型参数量降低了 79%，检测速度提升了 3.93 倍。与当前主流算法对比，也存在着一定的优势。但本文算法模型仍存在一定的改进空间，如何在不降低检测精度的条件下更好地提升检测速度是下一步的研究重点。

参考文献

[1] Chavez, S., Long, B., Koyfman, A. and Stephen, Y. (2020) Coronavirus Disease (COVID-19): A Primer for Emergency Physicians. *American Journal of Emergency Medicine*, **44**, 220-229. https://doi.org/10.1016/j.ajem.2020.03.036

[2] Tian, Z., Shen, C. and Chen, H. (2020) FCOS: Fully Convolutional One-Stage Object Detection. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 9627-9636. https://doi.org/10.1109/ICCV.2019.00972

[3] Chen, P., Li, Y. and Zhou, H. (2020) Detection of Small Ship Objects Using Anchor Boxes Cluster and Feature Pyramid Network Model for SAR Imagery. *Journal of Marine Science and Engineering*, **8**, 112. https://doi.org/10.3390/jmse8020112

[4] 张路达, 邓超. 多尺度融合的 YOLOv3 人群口罩佩戴检测方法[J]. 计算机工程与应用, 2021, 57(16): 283-290.

[5] Redmon, J. and Farhadi, A. (2018) YOLOv3: An Incremental Improvement.

[6] 李雨阳, 沈记全, 翟海霞, 冯伟华. 基于改进 SSD 的口罩佩戴检测算法[J/OL]. 计算机工程, 1-9. https://doi.org/10.19678/j.issn.1000-3428.0062150, 2022-06-29.

[7] Liu, W., Anguelov, D. and Erhan, D. (2016) SSD: Single Shot MultiBox Detector. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2

[8] 王艺皓, 丁洪伟, 李波, 杨志军, 杨俊东. 复杂场景下基于改进 YOLOv3 的口罩佩戴检测算法[J]. 计算机工程, 2020, 46(11): 11.

[9] He, K., Zhang, X. and Ren, S. (2016) Deep Residual Learning for Image Recognition. https://doi.org/10.1109/CVPR.2016.90

[10] Gong, H., Li, H. and Xu, K. (2019) Object Detection Based on Improved YOLOv3-Tiny. 2019 *Chinese Automation Congress* (*CAC*), Hangzhou, 22-24 November 2019, 3240-3245. https://doi.org/10.1109/CAC48633.2019.8996750

[11] Chollet, F. (2017) Xception: Deep Learning with Depthwise Separable Convolutions. 2017 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, 21-26 July 2017, 1251-1258. https://doi.org/10.1109/CVPR.2017.195

[12] Mahendran, A. and Vedaldi, A. (2016) Visualizing Deep Convolutional Neural Networks Using Natural Pre-Images. *International Journal of Computer Vision*, **120**, 233-255. https://doi.org/10.1007/s11263-016-0911-8

[13] Lin, M., Chen, Q. and Yan, S. (2013) Network in Network. Computer Science.

[14] Han, S., Pool, J., Tran, J. and William, J.D. (2015) Learning Both Weights and Connections for Efficient Neural Networks. *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Volume 1, 1135-1143.

[15] Geoffrey, H., Oriol, V. and Jeff, D. (2015) Distilling the Knowledge in a Neural Network. *Computer Science*, **14**, 38-39.

[16] Mingxing, T. and Quoc, V.L. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.

[17] Hu, J., Shen, L., Sun, G., Albanie, S. and Wu, E.H. (2017) Squeeze-and-Excitation Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **42**, 2011-2023.

[18] Wang, Q.L., *et al*. (2019) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. 2020 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, *CVPR* 2020, Seattle, June 2020, 13-19.