[文章编号] 1003－4684(2026)04-0001-06

基于改进 YOLOv5 的人脸口罩佩戴检测方法研究

郭 旋，周先军

（湖北工业大学电气与电子工程学院，湖北 武汉 430068）

[摘 要] 新冠疫情期间，正确佩戴口罩可以有效防止疫情扩散，传统的图像处理算法不能很好地对口罩佩戴进行检测，而经典的深度学习模型直接用于口罩佩戴的检测时精度低，收敛速度慢，效果不甚理想。针对上述问题，在 YOLOv5 模型的基础上，引入了 CBAM 注意力机制，解决模型的特征提取能力弱的问题；嵌入 C2f 结构，解决 C3 模块特征提取效率低的问题，进一步提高模型的检测精度和速度；使用 SIoU 损失函数替换 GIoU 损失函数，解决检测框定位不准确的问题。实验结果表明，与原始 YOLOv5s 模型相比，其 mAP 值提升了 2.3%，参数量减少了 14.13%，FPS 提升了 18.19%。

[关键词] 口罩检测；注意力机制；YOLOv5；损失函数

[中图分类号] TP39 [文献标识码] A

2020 年初，新冠疫情席卷全球，给国家和社会带来了巨大负担。由于病毒可以通过飞沫传播，理论上科学地佩戴口罩是防止疫情大规模传播最简单且方便的措施。《公众和重点职业人群戴口罩指引》指出："正确佩戴口罩，要确保口罩盖住口鼻。"即正确佩戴口罩，要求将口、鼻都覆盖在口罩之下。在现实生活中，许多人佩戴口罩都不规范，露出口鼻的佩戴方式随处可见，错误的佩戴方式降低了口罩的防护效果，使佩戴者自己和他人感染的风险都大幅增加。目前许多场所都只在入口处设立口罩检测，通过入口的佩戴检测之后，许多人都选择拉下口罩，露出鼻子进行呼吸。这种行为让佩戴口罩失去了意义，极大地增加了疫情传播的风险，尤其是在公交、电影院、电梯等人员密集场所。为避免因口罩佩戴不规范而带来的疫情传播风险，需要对口罩是否规范佩戴进行检测，因此检测活动人群是否正确佩戴口罩对于疫情防控具有非常重要的实际意义。

传统目标检测对目标进行特征提取，依靠的是手工设计的特征，该方法鲁棒性较差，而卷积神经网络（convolutional neural networks，CNN）能够提取对检测任务更加有用的特征，因此在复杂场景下，基于深度学习的目标检测算法的检测效果更佳。目前主流的算法有 SSD 算法[1]、RetinaNet[2]、R-CNN[3]、Fast R-CNN[4]、Mask R-CNN[5]、YOLO 系列等。YOLO 系列凭借其较为快速且准确的优点，被广泛应用于各个行业之中。

YOLO 系列算法最早由 Joseph Redmon 等[6]于 2015 年提出，之后 YOLO 系列不断推陈出新，开发出 YOLOv2[7]，YOLOv3[8]，YOLOv4[9]等多个版本，Ultralytics 公司于 2020 年 6 月 9 日提出了 YOLOv5，该算法基于 YOLOv3，使用了更多新兴技术，在速度与精度上都超过了前作。程长文等[10]在 YOLOv5 中加入自适应锚定框并改进了前馈输入层以优化数据处理，降低计算量，但是检测精度仍有较大提升空间。LOEY M 等[11]在 Backbone 卷积层中，融入了 ResNet-50 网络模型，用来检测医用口罩的佩戴情况，但是其准确率较低，对小型目标的检测效果较差。陈继平等[12]设计了一种高激活性鬼影（High Active Ghost，HAG）模块，减少了特征图冗余信息的同时，增加了模型的计算量。何义等[13]以 k-means 算法为基础，添加了聚类算法，获取与真实框之间的更高的先验框，在多人物以及实时检测场景中准确率有所提高，但是检测速度没有得到改善。

以上研究存在着一个共同的痛点：无法在提高检测速度的情况下，同时兼顾检测的精度。为解决上述问题，本文提出一种基于改进 YOLOv5 的口罩佩戴检测算法。主要改进如下：通过引入 CBAM 注意力机制，解决模型的特征提取能力弱的问题；嵌入 C2f 结构，解决 C3 模块特征提取效率低的问题，进

[收稿日期] 2023－07－25

[基金项目] 国家自然科学基金项目（61901165，61601177）；湖北省自然科学基金项目（2019CFB530）

[第一作者] 郭 旋（1996－），男，湖北武汉人，湖北工业大学硕士研究生，研究方向为深度学习、目标检测。

[通信作者] 周先军（1970－），男，湖北武汉人，湖北工业大学副教授，研究方向为人工智能、无线网络。

一步提高模型的检测精度和速度；使用 SIoU 损失函数替换 GIoU 损失函数，解决检测框定位不准确的问题。

1 基于 YOLOv5s 的口罩佩戴检测算法

YOLOV5s 在结构上分为 Backbone，Neck，Head 三个部分。Backbone 部分负责模型的特征提取，Neck 部分负责模型的特征融合，Head 部分负责对目标类别、置信度等进行预测。YOLOV5s 的模型结构如图 1 所示。

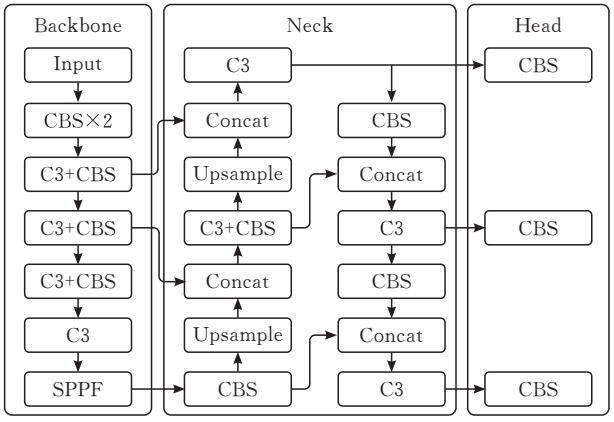


图 1 YOLOv5s 网络结构图

1.1 改进 YOLOv5s 网络模型

为了增强特征提取能力，在网络中插入 CBAM 注意力机制，同时为了提升检测速度和精度，将 C3 模块替换为 C2f 模块，最后使用 SIoU 损失函数替换原本的 GIoU 损失函数来优化边界框的回归过程，进一步提高检测模型的精度。改进后的网络模型如图 2 所示。

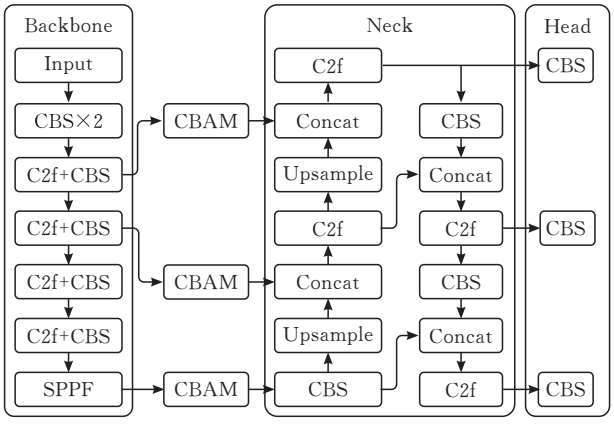


图 2 改进 YOLOv5s 网络结构图

1.1.1 轻量级注意力模块(CBAM，Convolutional Block Attention Module) 注意力机制的本质是提取出适用于该模型的一组注意力权重系数，简单来说就是在众多信息中提取出对当前任务目标更关键的信息，并且忽略或抑制除关键信息外的无关信息。2014 年 Volodymyr 等[14]最早将注意力机制应用在视觉领域，随后 Ashish Vaswani 等在 2017 年发表了《Attention is all you need》[15]一文，提出了 Transformer 结构，这种网络结构仅基于注意机制，完全舍弃了循环和卷积结构。在此之后，注意力机制在 NLP、CV 相关问题的网络设计上被广泛应用。

在深度学习中，注意力机制可以针对性提取相关性更大的部分，使神经网络能够提取出更需要的特征。CBAM[16]（convolutional block attention module）是一种轻量级注意力模块，它可以在空间维度和通道维度上进行 Attention 操作。CBAM 模块由通道注意力模块(Channel Attention Module，CAM)和空间注意力模块(Spatial Attention Module，SAM)2 个子模块组成，CAM 模块和 SAM 模块分别进行通道与空间上的 Attention。这种设计在能够减少参数和节约计算力的同时，保证了 CBAM 模块可以较为方便地添加到目标网络结构中，其网络结构如图 3 所示。

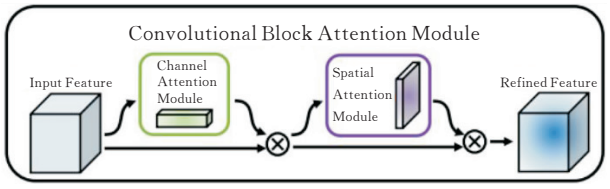


图 3 CBAM 网络结构图

CAM 模块关注的是图像是什么，即图像的类别，具体流程为：首先分别对输入图像进行最大池化和平均池化，对特征图进行维度压缩，再分别送入一个两层的神经网络，共享全连接，将两个特征图基于元素对应相乘的加和操作后，再经过 sigmoid 函数进行激活，得到含有权重的通道注意力特征图。通道注意力机制的计算公式如下：

$$F = \frac{1}{C}\sum_{i\in C}F(i) + \max_{i\in C}F(i) \tag{1}$$

$$M_C(F) = \sigma(W_1(W_2(F^c_{avg})) + W_1(W_2(F^c_{\max}))) \tag{2}$$

C 代表输入特征图的通道数；F^c_{avg} 和 $F^c_{\max}$ 分别为经过平均值池化和最大值池化的特征图；W_1 和 W_2 为 MLP 的权重。

SAM 模块关注的是图像在哪里，即图像的位置，具体流程为：首先对特征图进行最大池化和平均池化处理，基于通道进行压缩，沿通道汇聚空间维度的信息，然后将得到的特征图进行 concat 拼接，再经过卷积进行降维，最后经过 Sigmoid 激活操作得到空间注意力特征。其计算公式如下：

$$\begin{aligned} M_S(F) &= \sigma(f^{7\times7}([AvgPool(F);MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^s_{avg};F^s_{\max}])) \end{aligned} \tag{3}$$

1.1.2 C2f 模块 C2f 模块来源于 C3 模块和

ELAN[17]模块,C3 模块如图 4 所示,C2f 模块采用了更加轻量化的设计方案,如图 5 所示。该模块包括 Bottleneck 模块和 CBS 模块两个子模块,其中 Bottleneck 模块是由两个 CBS 模块堆叠而来。CBS 模块由 Conv2d、BatchNorm2d 和 SiLU 激活函数堆叠而成,上述模块可以高效地提取特征,减少计算量和模型大小。

除了卷积层,C2f 模块还采用了残差连接与跳跃连接来提高模型性能,使得网络更好地学习到特征,残差连接能够避免梯度消失和梯度爆炸等问题。跳跃连接则可以提高特征的语义表达能力,从而增强网络的特征提取能力。上述结构可以帮助网络更好地学习到特征。从而帮助网络更好地进行特征提取和分类,进而提高检测精度和速度。此外,C2f 模块还可以提高网络的鲁棒性,使得网络在处理各种不同的场景和图像时都能够表现出良好的性能。

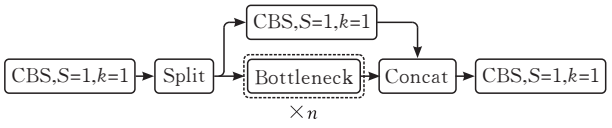


图 4　C3 结构图

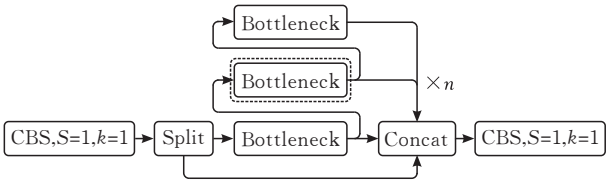


图 5　C2f 结构图

1.1.3　SIoU 损失函数　损失函数是用来度量模型预测值与真实值差异程度的运算函数,损失函数的定义可以在较大程度上影响检测效果。传统的目标检测损失函数,如 YOLOv5 中包含的 GIoU[18]、CIoU[19]等,关注的是预测框和真实框的距离、重叠区域和纵横比等指标,而这些方法都没有考虑真实框与预测框之间的角度关系,从而导致检测速度慢,检测结果不准确等问题。SIoU 损失函数考虑到真实框与预测框之间的角度关系,添加了角度惩罚成本,提高了检测速度和检测结果的准确度,SIoU 损失函数示意图如图 6 所示。

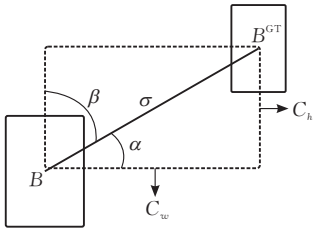


图 6　SIoU 示意图

图中 B 代表预测框;B^{GT} 代表真实框;C_w 和 C_h 分别为二者中心点的横向距离和纵向距离;α 和 β 为二者中心点连线与水平线和垂直线的夹角;σ 为二者中心点的距离。

SIoU 损失函数由角度成本、距离成本、形状成本和 IoU 成本 4 个成本函数组成。角度成本的计算公式为:

$$x = \frac{c_h}{\sigma} = \sin(\alpha) \tag{4}$$

$$d = \sqrt{(b_{cx}^{gt} - b_{cx})^2 + (b_{cy}^{gt} - b_{cy})^2} \tag{5}$$

$$C_h = \max(b_{cy}^{gt}, b_{cy}) - \min(b_{cy}^{gt}, b_{cy}) \tag{6}$$

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{7}$$

式中:Λ 是角度计算的最终结果;x 是角度 α 的正弦值;d 是预测框和真实框之间的距离;C_h 是二者中心点间的相对高度差。

距离成本计算公式如下:

$$\rho_x = \left(\frac{b_{cx}^{gt} - b_{cx}}{c_w}\right)^2, \rho_y = \left(\frac{b_{cy}^{gt} - b_{cy}}{c_h}\right)^2, \gamma = 2 - \Lambda \tag{8}$$

$$\Delta = \sum_{t=x,y} (1 - e^{-\gamma\rho^t}) \tag{9}$$

式中:ρ_x 和 ρ_y 是真实框和预测框的中心点在 x 轴和 y 轴上的相对距离与它们的最小外接矩形的宽高的比值的平方;e 为欧拉常数。

形状成本的计算公式如下:

$$\omega_w = \frac{|w - w^{gt}|}{\max(w, w^{gt})}, \omega_h = \frac{|h - h^{gt}|}{\max(h, h^{gt})} \tag{10}$$

$$\Omega = \sum_{t=w,h} (1 - e^{-\omega_t})^{\theta} \tag{11}$$

式中:(w, w^{gt}) 和 (h, h^{gt}) 分别是预测框和真实框的宽高;Ω 为形状计算公式中的注意力系数。

最后是 IoU 成本的计算方法,即 SIoU 的最终表达式,其公式如下:

$$\mathrm{IoU} = \frac{|B \cap B^{GT}|}{|B \cup B^{GT}|} \tag{12}$$

$$\mathrm{SIoU} = 1 - \mathrm{IoU} + \frac{\Delta + \Omega}{2} \tag{13}$$

2　实验过程

2.1　数据集介绍

由于人脸口罩佩戴检测的公开数据集没有佩戴错误这一分类,本文采用自制的人脸口罩佩戴检测数据集,检测的内容包括未佩戴口罩(no_mask)、佩戴口罩(have_mask)和错误佩戴口罩(incorrect_mask)三类。经过筛选共获得 12351 张图片,其中训练集图片占比 70%,验证集图片有 1885 张,测试集图片有 1885 张。本文使用了 mosaic 数据增强技术,其主要操作是将四张图片随机进行裁剪和拼接,得到一张新的图片,然后就这张新的图片作为训练数据,从而丰富检测图像的背景和小目标,提高网络模型的训练效果。



图 7 数据集样张



图 8 mosaic 数据增强处理结果

2.2 实验评估指标

本文使用平均精度均值(mean average precision,mAP)和每秒传输帧数(FPS)作为实验的主要评估指标,mAP 的计算公式如下:

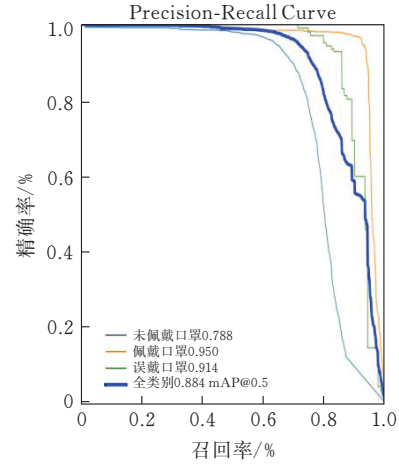
$$mAP = \frac{\sum_{i=l}^{k} AP_i}{k} \tag{14}$$

其中,k 表示类别数;AP_i 表示第 i 类的 PR 曲线下面积。

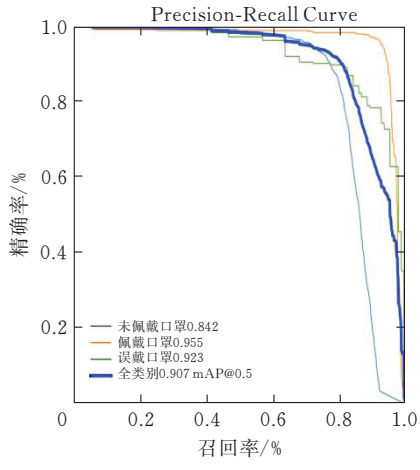
2.3 实验环境及结果

实验采用深度学习框架 PyTorch 训练模型,操作系统 Windows 10 专业版,编程工具:pycharm。硬件环境如下:13th Gen Intel(R) Core(TM) i9-13900K 处理器、Nvidia GeForce RTX 4080、32GB RAM。初始学习率为 0.01,batch_size 为 64,总 epoch 为 100,网络输入尺寸为 640×640,IoU 阈值设置为 0.5。

YOLOv5 算法改进前后的 mAP 对比曲线如图 9 所示。由图可见,原始 YOLOv5 模型在未佩戴口罩、佩戴口罩、误戴口罩以及全类别上的 mAP 值分别为 0.788、0.950、0.914 和 0.884,而改进后的模型在未佩戴口罩、佩戴口罩、误戴口罩以及全类别上的 mAP 值分别为 0.842、0.955、0.923 和 0.907,在所有类别上,改进后模型的 mAP 值均有明显提升。



(a) YOLOv5s



(b)改进 YOLOv5s

图 9 改进前后 mAP 曲线

2.4 消融实验和结果分析

为了验证改进方法的有效性和泛化性,消融实验分别对 YOLOv5s 模型以及本算法所提出修改的模型做出对比验证。消融实验如表 1 所示,表中"√"代表使用该策略,"×"代表未使用该策略。如表 1 所示,各个模型由第一列的序号表示:①代表了 YOLOv5s 模型;②代表了将原有 C3 模块更换为 C2f 模块;③代表了在②的基础上使用 SIoU 损失函数;④代表在③的基础上添加了 CBAM 注意力机制。

表 1 消融实验

模型	CBAM	C2f	SIoU	mAP^{all}/%	FPS
①	×	×	×	0.884	384.62
②	√	×	×	0.89	400
③	√	√	×	0.903	400
④	√	√	√	0.907	454.55

由表 1 可知,对比模型①和②的数据可知,模型②的平均精度均值提升了 0.6 个百分点。在小幅提升检测精度的同时,模型的 FPS 也有一定的提升。模型②将 YOLOv5 原本网络结构中的 C3 模块替换

为 C2f 模块，C2f 模块能提高特征的语义表达能力，从而提高模型的检测精度和速度。

对比模型②和③的数据可知，模型③的平均精度均值提升了 1.3 个百分点，模型③在模型②的基础上，将 YOLOv5 中原本的损失函数 GIoU 替换为 SIoU，SIoU 能提高口罩和人脸边框定位精度，从而提升模型的检测精度。

对比模型③和④的数据可知，模型④的平均精度均值提升了 0.4 个百分点，模型的 FPS 提升了 13.63%，模型④在模型③的基础上，在 YOLOv5 中加入了 CBAM 注意力机制，能有效提高模型的特征提取能力，从而提高模型的检测精度和速度。

2.5　对比实验和结果分析

将本文模型与其他主流算法相比较，来检验本文模型的性能以及优势，实验结果如表 2 所示。

表 2　对比试验

模型	mAP@0.5/%	FPS	Params/10^6	GFLOPs	模型大小/MB
YOLOv5s	0.884	384.62	7.02	15.8	14.03
YOLOv5s-GhostNet	0.875	476.19	3.68	8	7.58
YOLOv3	0.9	119.05	61.51	154.6	120.58
YOLOv3-tiny	0.864	476.19	8.67	12.9	17.02
Ours	0.907	454.55	6.03	13.7	12.07

由表 2 可知，与原始 YOLOv5s 模型进行对比，本文提出的模型 mAP 提升 2.3%，参数量减少 14.13%，FPS 提升 18.19%；与 YOLOv3 这种大模型进行对比，模型大小减少了 90%，且 mAP 值仍有 0.7% 的提升；与 YOLOv5s-GhostNet 这种小模型进行对比，mAP 值增加了 3.2%，FPS 仅减少了 4.5%。

不同模型检测效果对比图如图 10 所示，本文提出的改进模型在各个类别的检测精度都有较好表现。



(a) YOLOv5s-GhostNet　(b) YOLOv3

(c) YOLOv5s　(d) ours

图 10　不同模型检测效果对比图

3　结束语

为了提升人脸口罩佩戴检测任务的检测速度和精度，本文提出了基于改进的 YOLOv5 口罩佩戴检测模型，引入 CBAM 注意力机制，提高了模型的特征提取能力；使用 C2f 结构可提高特征的语义表达能力，增强网络的特征提取能力，从而有利于模型更准确地检测小目标；最后将 GIoU 损失函数替换 SIoU 为损失函数，获得更准确的锚框。与原始 YOLOv5s 模型相比，本文方法的平均精度均值提升 2.3%，参数量减少 14.13%，FPS 提升 18.19%，证明了本文方法对口罩佩戴检测具有较好的效果。与当下主流的检测算法相比，精度和速度均有不同程度提升，但在检测精度上仍存在一定的改进空间，在之后的研究中会在保证检测速度的前提下，进一步关注检测精度。

[参　考　文　献]

[1] Liu W, Anguelov D, Erhan D, et al. Ssd: single shot multibox detector[C]// Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

[2] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.

[3] Girshick R, Donahue J, Darrell T ,et al.Rich feature hierarchies for accurate object detection and semantic segmentation[J].IEEE Computer Society, 2014.DOI: 10.1109/CVPR.2014.81.

[4] Girshick R. Fast R-CNN[J].Computer Science, 2015. DOI:10.1109/ICCV.2015.169.

[5] He K, Gkioxari G, Piotr Dollár, et al. Mask R-CNN [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.DOI: 10.1109/TPAMI.2018.2844175.

[6] 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. 2016: 779-788.

[7] Edmon J, Farhadi A. YOLO9000: better, faster, stronger[C] // 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 7263-7271.

[8] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2021-11-01]. https: //arxiv. org/abs/1804. 02767.

[9] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[10] 程长文,陈玮,陈劲宏,等.改进 YOLO 的口罩佩戴实时检测方法[J].电子科技, 2023, 36(02):8.

[11] Loey M, Manogaran G, Taha M H N, et al. Fighting against COVID-19: A novel deep learning model based on YOLO-v2 with ResNet-50 for medical face mask detection[J]. Sustainable cities and society, 2021, 65: 102600.

[12] 陈继平,陈永平,谢懿,等.Ghost-YOLO:轻量化口罩人脸检测算法[J].信号处理,2022,38(09):1954-1964.

[13] 何义,李捍东.基于改进 YOLOv5 模型的复杂场景口罩佩戴识别[J].微处理机, 2022(002):043.

[14] Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. Advances in neural information processing systems, 2014,27:2204-2212.

[15] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017,30:5998-6008.

[16] Woo S, Park J, Lee J Y, et al. Cbam: convolutional block attention module[C]// Proceedings of the European conference on computer vision (ECCV), 2018: 3-19.

[17] Zhang X, Zeng H, Guo S, et al. Efficient long-range attention network for image super-resolution[C] // Computer Vision-ECCV 2002: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XⅦ. Cham: Springer Nature Switzerland, 2022: 649-667.

[18] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: a metric and a loss for bounding box regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019. DOI:10.1109/CVPR.2019.00075.

[19] Chen D , Miao D .Control distance iou and control distance IoU loss function for better bounding box regression[J]. Pattern Recognition, 2021. DOI: 10. 48550/arXiv.2103.11696.

Research on Face Mask Wearing Detection Method Based on Improved YOLOv5

GUO Xuan, ZHOU Xianjun

(*School of Electrical and Electronic Engineering, Hubei Univ. of Tech., Wuhan* 430068, *China*)

Abstract: During the COVID-19 epidemic, proper mask wearing can effectively prevent the spread of the epidemic. Traditional image processing algorithms cannot detect mask wearing well, while the classical deep learning model is less effective when directly used for the detection of mask wearing with low detection accuracy and slow convergence speed. To address the above problems, this paper, on the basis of the YOLOv5 model, solves the problem of weak feature extraction ability of the model by introducing the CBAM attention mechanism; embeds the C2f structure to solve the problem of low feature extraction efficiency of the C3 module and further improves the detection accuracy and speed of the model; uses the SIoU loss function to replace the GIoU loss function to solve the problem of inaccurate detection frame localization. The experimental results show that compared with the original YOLOv5s model, the mAP of this paper's method is improved by 2.3%, the parameters are reduced by 14.13% and the FPS is improved by 18.19%, which proves that this paper's method has better effect on mask wearing detection.

Keywords: mask detection; attention mechanism; YOLOv5; loss function

[责任编校: 赵 瞳]