基于改进YOLOv5的小目标检测算法

郭磊[1]，王邱龙[2]，薛伟[2]，郭济[3]

1.电子科技大学，计算机科学与工程学院 成都 611731

2.新疆大学，信息科学与工程学院 乌鲁木齐 830000

3.西藏民族大学，财经学院 咸阳 712082

【摘要】为解决目标检测中对小目标误检、漏检及特征提取能力不足等问题，提出基于改进YOLOv5的小目标检测算法。算法使用Mosaic-8方法进行数据增强，通过增加一个浅层特征图，调整损失函数，增强网络对小目标的感知能力；通过修改目标框回归公式，解决训练过程中梯度消失等问题，提升了小目标检测精度。将改进后的算法应用在密集人群情景下的防护面具佩戴检测中，实验结果表明，相较于原始YOLOv5算法，该算法在小目标检测上具有更强的特征提取能力和更高的检测精度。

关 键 词 数据增强; 深度学习; 小目标检测; YOLOv5

中图分类号 TP39 文献标识码 A

A Small Object Detection Algorithm based on Improved YOLOv5

GUO Lei[1], Wang Qiulong[2], XUE Wei[2], and GUO Ji[3]

1. School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731

2. School of Information Science and Engineering, Xinjiang University Urumqi 830000

3. College of Finance and Economics, Xizang Minzu University Xianyang 712082

Abstract For object detection, one immediate problem is the insufficiency of feature extraction on small objects, which is easy to make false detection and miss the inspection on small targets. To solve the problem of small object detection, an improved detection algorithm based on YOLOv5 was proposed. The algorithm uses the method of Mosaic-8 on data augmentation. A shallow feature map is added to the YOLOv5 network and loss function is adjusted to improve the sensibility of network on small targets. The target box regression formula is modified to solve the problem of gradient disappearance in training process, which realized accurate precision on small targets. The improved algorithm is applied to mask wearing detection under crowed environment. Experimental results show that the proposal algorithm has stronger feature extraction ability and higher detection accuracy on small object detection compared to the original YOLOv5 algorithm.

Key words Data Augmentation; Deep Learning; Small Object Detection; YOLOv5

1 引言

目标检测作为计算机视觉领域的核心问题之一，其任务就是找出图像中所有感兴趣的目标，并确定它们的位置和类别。目标检测结合了目标定位和目标分类两大任务，被广泛应用于人脸识别[1] 、自动驾驶[2 - 3] 、行人检测[4] 、智能监控[5] 等计算机视觉领域。作为图像理解、视频理解的基石，目标检测是解决图片分割、目标跟踪、图像描述、事件检测和场景理解等更高层次视觉任务的基础。

传统的目标检测算法由3部分组成，分别是区域选择、特征提取和分类器，但由于其存在手工设计的特征鲁棒性差、区域选择策略没有针对性等特点，检测效果并不理想。随着深度学习的深入发展，基

收稿日期： – – ；修回日期： – –

基金项目：国家重点研发计划项目“多源涉诉信访智能处置技术研究”(2018YFC0831800)

作者简介：郭磊(1971 –)，男，博士，副教授，硕士生导师，主要从事机器学习、视频理解、嵌入式系统方面的研究；王邱龙(1997-)，男，硕士研究生，主要研究方向为计算机视觉、机器学习；薛伟(1997-)，男，硕士研究生，主要研究方向为计算机视觉、算法、机器学习；郭济(2002-)，男，本科，主要研究方向为统计学。

于深度卷积神经网络的特征提取技术被广泛应用到计算机视觉任务中，目标检测完成了从基于传统手工设计特征的检测方法到基于卷积神经网络的深度学习方法的变迁，随后基于卷积神经网络的目标检测算法迅速成为图像处理领域研究的主流[6] 。目前两个令人瞩目的方法是以R-CNN(Region-based Convolutional Neural Network)系列为代表的基于候选框的两阶段深度学习算法和以YOLO(You Only Look Once)系列为代表的基于回归的单阶段深度学习目标检测算法。

小目标检测长期以来就是目标检测中的重点和难点之一。目标检测中对小目标的定义通常有两种，第一种是相对尺寸大小，根据国际光程学会定义，小目标为256×256像素的图像中成像面积小于80个像素点的目标，即若图像中目标的尺寸小于原图的0.12%则可以认为是小目标。另一种是根据具体的数据集对小目标进行定义，例如在COCO数据集[7] 中，尺寸小于32×32像素的目标被认为是小目标；Zhu等人[8] 在其交通标志数据集中，将宽度占整个图像比例小于20%的目标定义为小目标；而数据集CamVid中则将类别sign symbol, pedestrian pole和bicycle定义为小目标，其他类别定义为常规目标。如图 1所示，图 1(a)中区域标注的三个对象表示常规尺寸，图1(b)中区域标注的三个对象表示小目标。小目标相比于常规目标，在图像中所占的像素数较少，分辨率低，信息量少，特征表达能力弱。

(a) 常规尺寸目标 (b) 小目标

图 1 常规目标与小目标示例

2 相关工作

早期的目标检测方法是使用手工提取特征，再在此基础上构造模型。使用该方法设计的模型不仅结构复杂，而且难以提升精度。随着深度学习的发展，人们发现卷积神经网络具有极好的学习特征的能力，因此，Girshck等人[9] 将CNN应用到目标检测中，提出了R-CNN模型，使得模型的检测性能有了很大的提升，但其网络只能输入固定尺寸的图像，且R-CNN通过选择性搜索(SS, Selective Search)算法生成的候选区域大量重叠，造成了计算资源的极大浪费。针对R-CNN模型的这一缺点，He等人[10] 提出了SPP-Net网络，Grishick与Ren等人[11-12] 先后提出了Fast R-CNN和Faster R-CNN模型。从Fast R-CNN模型发展到Faster R-CNN模型，虽然模型的检测速度和精度均有所提高，但是这类两阶段目标检测算法与单阶段目标检测算法在检测速度上相比仍具有一定的差距。

经典的单阶段目标检测算法有YOLO系列算法和SSD(Single Shot MultiBox Detector)算法。Redmon等人[13] 在2015年提出了第一个单阶段目标检测算法YOLO，Liu等人[14] 借鉴YOLO和Faster R-CNN的优点提出了SSD算法，提高了检测速度并实现了多尺度预测。Redmon等人[15 -16] 在YOLOv1的基础上继续改进，又提出了YOLOv2和YOLOv3检测算法，其中YOLOv2进行了多种尝试，使用了批标准化(BN, Batch Normalization)技术,引入了锚框机制；YOLOv3采用darknet-53作为骨干网络，并且使用了三种不同大小的锚框，在逻辑分类器中使用sigmoid函数把输出约束在0到1之间，使得YOLOv3拥有更快的推理速度。Alexey等人[17] 于2020年提出了YOLOv4算法，在传统的YOLO基础上，加入了一些实用的技巧，将BackBone骨干网络中的ReLU激活函数改为了Mish激活函数，与ReLU相比，Mish函数图像更加平滑，实现了检测速度和精度的最佳权衡，使得其可以在一块普通的GPU(1080Ti)上完成网络训练。从YOLOv1至今，YOLO系列已经发展到了YOLOv5，相较于之前的YOLO版本，YOLOv5的网络结构更加灵活，在某中程度上YOLOv5已经成为YOLO系列算法中的SOTA(State Of The Art)。

YOLOv5是一个高性能的通用的目标检测模型，能一次性完成目标定位与目标分类两个任务，因此选择YOLOv5作为目标检测的基本骨架是可行的。但是为了实现一些场景下对小目标的独特性检测，就需要对YOLOv5的网络结构进行相应的调整和改进。

3 改进YOLOv5算法

本文在YOLOv5网络的基础上进行改进，改进后的整体的网络结构如图 2所示，黑色虚线框表示对原始网络的修改部分，通过新增尺寸为输入图像尺寸四分之一的特征图来提升对小目标特征的挖掘，

采用多尺度反馈以引入全局上下文信息来提升对图像中小目标的识别能力。损失函数使用CIOU[18]，从重叠面积、中心点距离、长宽比三个方面更好的描述目标框的回归。在原始YOLOv5的基础上使用Mosaic-8数据增强，修改目标框回归的公式，提高模型的收敛精度。下面分别从Mosaic-8数据增强、特征提取器、损失函数和目标框回归四方面进行详细介绍。

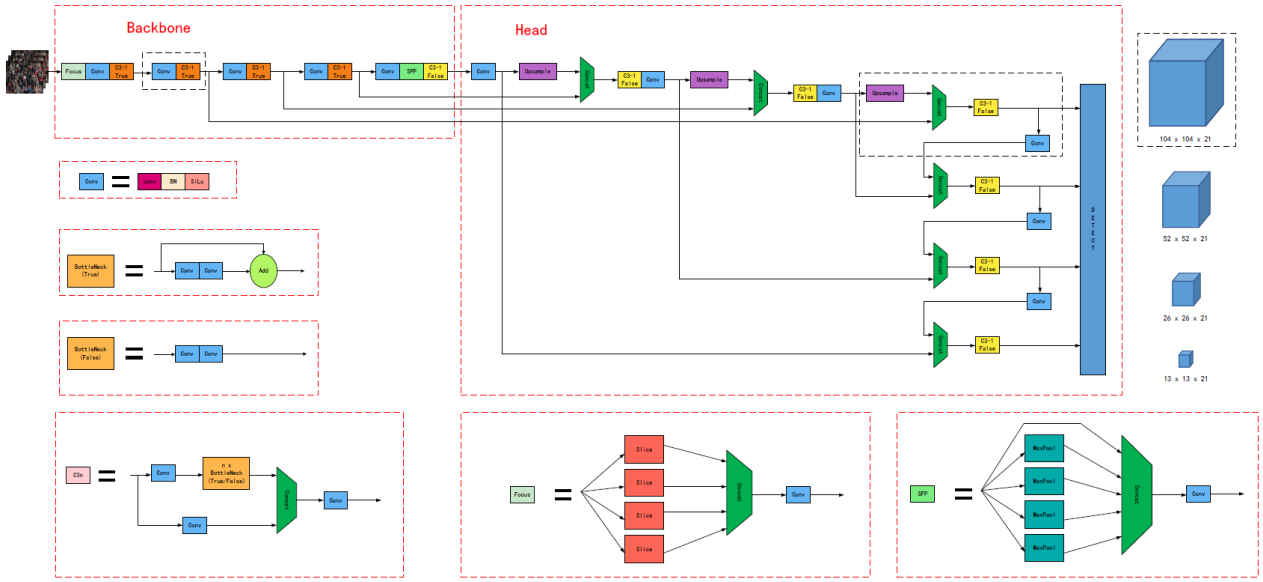


图 2 整体网络结构图

3.1 Mosaic-8数据增强

一个优秀的神经网络，往往需要大量的数据作支撑，然而获取新的数据这项工作往往需要花费大量的时间与人工成本，因此数据增强应运而生，它可以充分利用计算机来生成数据，增加数据量，例如采用缩放、平移、旋转、色彩变换等方法增强数据，数据增强的好处是能够增加训练的样本数量，同时噪声数据也相应增多，能够提高模型的泛化力。

在YOLOv5中除了使用最基本的数据增强方法外，还是用了Mosaic数据增强方法，其主要思想就是将四张图片，进行随机裁剪、缩放后再随机排列拼接形成一张图片，实现丰富数据集的同时，增加了小样本目标，提升网络的训练速度。在进行归一化操作的时，会一次性计算四张图片，小批量则不需要很大，因此模型对内存的需求降低。Mosaic数据增强的流程如图 3所示。

本文受Mosaic思想的启发，采用Mosaic方法的增强版——Mosaic-8，即采用8张图片随机裁剪、随机排列、随机缩放，然后组合成一张图片，以此来增加样本的数据量，同时合理引入一些随机噪声，增强网络模型对图像中小目标样本的区分力，提升模型的泛化力，其细节如图 4所示。

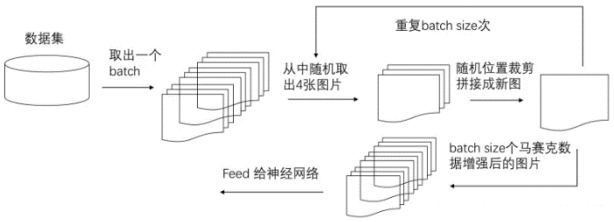


图 3 Mosaic 数据增强流程

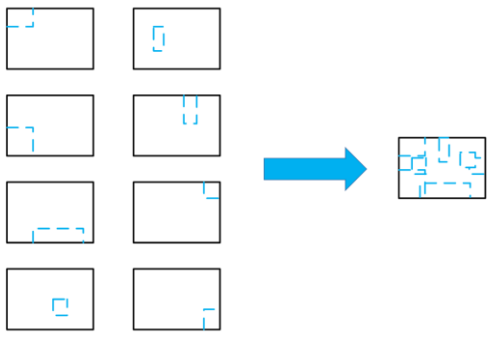


图 4 Mosaic-8 数据增强细节

3.2 特征提取器

在原始YOLOv5骨干网络中，使用3种不同尺寸的特征图来检测不同大小的目标，如图 5所示，该

网络将原始输入图像通过8倍下采样、16倍下采样、32倍下采样得到3中不同尺寸大小的特征图，将其输入到特征融合网络中。根据特征金字塔网络(FPN，Feature Pyramid Network) [19] 的思想可以看出，经过深层次卷积后的特征图虽然拥有丰富的语义信息，但在多次卷积的过程中会丢失掉目标的一些位置信息，不利于小目标的检测；而浅层卷积后得到的特征图语义信息虽然不够丰富，但是目标的位置信息却比较精确。

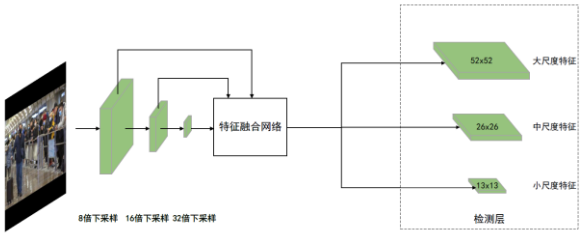


图 5 原始YOLOv5特征提取模型

在密集人群的条件下，大部分人脸检测目标占整幅图像的比例较小。因此，本文在YOLOv5骨干网络的基础上对原始输入图片增加一个4倍下采样的过程，如图 6所示。原始图片经过4倍下采样后送入到特征融合网络得到新尺寸的特征图，该特征图感受野较小，位置信息相对丰富，可以提升检测小尺寸口罩佩戴目标的检测效果。

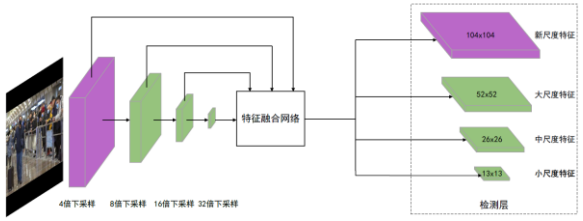


图 6 改进后的特征提取模型

在卷积神经网络中，经过不同的卷积层的得到的特征图含有目标不同的特征信息。浅层卷积后得到的特征图分辨率较高，目标位置信息相对丰富，但语义信息不明显；深层卷积后得到的特征图分辨率低，语义信息丰富，但丢失了较多的目标位置信息。因此，浅层特征图能区分较为简单的目标，深层特征图能区分复杂的目标，将低层特征图与高层特征图进行信息融合更有利于目标的区分。如图 7所示，将特征金字塔网络与路径聚合网络(PAN, Path Aggregation Network)[20] 相结合，特征金字塔网络自顶向下传递深层次语义特征，路径聚合网络自底向上传递目标的位置信息，通过顶向下和自底向上的特征信息融合有利于模型更好的学习到特征，增强模型对小目标和遮挡目标的敏感度。

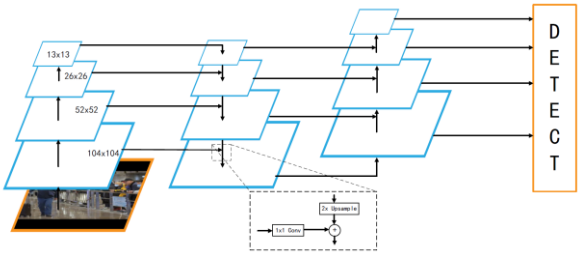


图 7 改进后的特征融合网络

3.3 损失函数

原始YOLOv5损失函数如公式 (1) 所示，它由定位损失、置信度损失和类别损失三部分构成。其中置信度损失和类别损失采用二元交叉熵损失函数进行计算，分别如公式 (3) 和公式 (4) 所示。其中，K 表示网络最后输出的的特征图划分为 $K\times K$ 个格子，M 表示每个格子对应的锚框的个数，I_{ij}^{obj} 表示有目标的锚框，I_{ij}^{noobj} 表示没有目标的锚框，λ_{noobj} 表示没有目标锚框的置信度损失权重系数。

$$Loss_{object} = Loss_{loc} + Loss_{conf} + Loss_{class} \tag{1}$$

$$Loss_{loc} = 1 - GIoU \tag{2}$$

$$\begin{aligned} Loss_{conf} = &-\sum_{i=0}^{K\times K} I_{ij}^{obj}[\hat{C}_i^j log(C_i^j) + (1-\hat{C}_i^j) log(1-C_i^j)] \\ &-\lambda_{noobj}\sum_{i=0}^{K\times K}\sum_{j=0}^{M} I_{ij}^{noobj}[\hat{C}_i^j log(C_i^j) + (1-\hat{C}_i^j) log(1-C_i^j)] \end{aligned} \tag{3}$$

$$Loss_{class} = -\sum_{i=0}^{K\times K} I_{ij}^{obj} \sum_{c\in classes} [\hat{P}_i^j log(P_i^j) + (1-\hat{P}_i^j) log(1-P_i^j)] \tag{4}$$

如图 8所示，黑色框是真实框，记为GT，绿色框是预测框，记为P，灰色框是同时包裹真实框和预测框的最小框，记为C。其中c是灰色框对角线的长度，d 是真实框中心点与预测框中心点的长度。

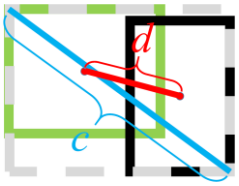


图 8 预测框P与真实框GT

原始YOLOv5 中使用GIoU[21] 来计算定位损失，如公式 (2) 所示，GIoU的计算如公式 (5) 所示。

$$GIoU = IoU - \frac{|C - GT \cup P|}{|C|} = \frac{|P \cap GT|}{|P \cup GT|} - \frac{|C - GT \cup P|}{|C|} \tag{5}$$

与原始IoU不同，GIoU不仅关注真实框与预测框之

间的重叠面积，还关注其他的非重叠区域，因此GIoU相较于原始IoU能更好的反应两者之间的重合度，但GIoU始终只考虑真实框与预测框之间的重叠率这一个因素，不能很好的描述目标框的回归问题。如图 9所示，当预测框在真实框内部时，且预测框的大小相同时，此时GIoU会退化为IoU，无法区分各个预测框之间的位置关系。

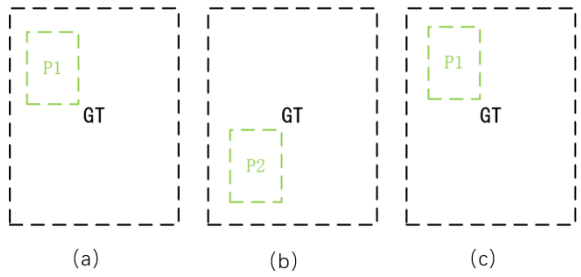


图 9 GIoU退化为IoU示例

本文选择CIoU替代GIoU作为目标框回归的损失函数，CIoU的计算如公式 (6) 所示。其中 α 是一个平衡参数，不参与梯度计算，α 的定义如公式 (7) 所示，v 是用来衡量长宽比一致性的参数，v 的定义如公式 (8) 所示。

$$CIoU = IoU - \frac{\rho^2\left(\mathbf{b},\mathbf{b}^{gt}\right)}{c^2} - \alpha v \tag{6}$$

$$\alpha = \frac{v}{(1-IoU)+v} \tag{7}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \tag{8}$$

CIoU综合考虑了真实框与预测框之间的重叠率、中心点距离、长宽比，能使得目标框回归更加稳定，收敛的精度更高。

3.4 目标框回归

目标框回归的目的就是要寻找某种映射关系，使得候选目标框(Region Proposal)的映射无限接近于真实目标框(Ground Truth)。对真实目标框的预测，通过预测相对位置的方法预测出目标框相对于坐上角的相对坐标。先验框与预测框的关系如图 10所示，其中，虚线框表示先验框，实线框表示预测框。预测框通过先验框平移缩放得到。将原始图片根据特征图尺寸划分成 $S\times S$ 个网格单元，每个网格单元会预测3个预测框，每个预测框包含4个坐标信息和1个置信度信息。当真实框中某个目标中心坐标落在某个网格中时，就由该网格预测这个目标。

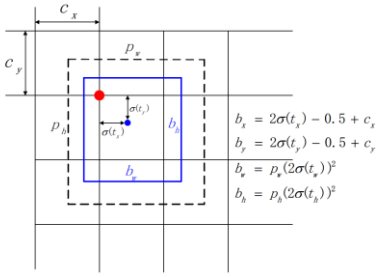


图 10 目标框回归原理图

目标框的坐标预测计算公式如下:

$$b_x = 2\sigma\left(t_x\right) - 0.5 + c_x \tag{9}$$

$$b_y = 2\sigma\left(t_y\right) - 0.5 + c_y \tag{10}$$

$$b_w = p_w\left(2\sigma\left(t_w\right)\right)^2 \tag{11}$$

$$b_h = p_h\left(2\sigma\left(t_h\right)\right)^2 \tag{12}$$

$$P_r(object) * IOU(b,object) = \sigma\left(t_o\right) \tag{13}$$

网络模型预测得到4个偏移 t_x 、t_y 、t_w、t_h，σ 表示Sigmoid激活函数，用于将网络预测值 t_x 、t_y 、t_w、t_h 映射到[0,1]之间，c_x 、c_y 单元网格中相对于图片左上角的偏移量，p_w 、p_h 是先验框宽高。通过上述公式最终得到预测目标框的中心坐标 b_x 、b_y 和宽高 b_w 、b_h 。$\sigma(t_o)$ 是预测框的置信度，由预测框的概率和预测框与真实框的IoU值相乘得到。对 $\sigma(t_o)$ 设定阈值，过滤掉置信度较低的预测框，然后再对剩下的预测框用非极大值抑制算法(NMS，Non-Maximum Suppression)[22] 得到最终的预测框。

在最小的特征图上，由于其感受野最大，故应该用其来检测大目标，所以大尺度的特征图应该应用小尺寸的先验框，小尺寸的特征图应该应用大尺度的特征图来进行预测框的回归。本文采用4尺度检测结构，4个尺度的特征图大小与先验框尺寸的对应关系如表 1所示。

表 1 特征图大小与先验框尺寸对应关系

特征图大小	先验框尺寸		
13×13	[116,90]	[156,198]	[373,326]
26×26	[30,61]	[62,45]	[59,119]
52×52	[10,13]	[16,30]	[33,23]
104×104	[5,6]	[8,14]	[15,11]

4 实验与结果分析

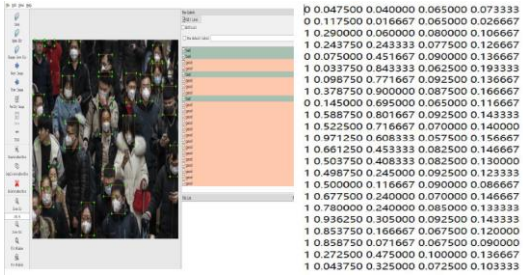
由于密集人群条件下往往人物众多，且容易出现人与人之间相互遮挡的现象，检测难度大，且单个人员的口罩占整幅图像的比例远远小于20%，因此可以将其作为小目标对待。将改进后的算法应用在密集人群的口罩佩戴场景下，并与参考文献[23]提出的算法、AIZOO算法和原始YOLOv5算法进行对比实验。

4.1 数据集

本文数据集来源于WIDER FACE、MAPA(Masked Faces)这两个公开数据集和网络，从中手动筛选出含有多人场景下的佩戴口罩和未佩戴口罩的人脸图片，最终得到训练集4000张，测试集1320张，共计5320张。部分数据集图片如图 11所示。

图 11 数据集部分图片

利用标记软件LabelImg对数据集进行YOLO格式的标注，共有两个标记类别，分别是bad(未佩戴口罩)和good(佩戴口罩)。标注完成后，每一张图片都对应着一个与该图片名称相同的txt文件，txt文件中的每一行表示一个标记实例，共5列，从左到右分别表示标签类别、标记框中心横坐标与图片宽度的比值、标记框中心纵坐标与图片高度的比值、标记框宽度与图片宽度的比值、标记框高度与图片高度的比值。数据集标注示例如图 12所示。



```
0 0.047500 0.040000 0.065000 0.073333
0 0.117500 0.016667 0.065000 0.026667
1 0.290000 0.060000 0.080000 0.106667
1 0.243750 0.243333 0.077500 0.126667
0 0.075000 0.451667 0.090000 0.136667
1 0.033750 0.843333 0.062500 0.193333
1 0.098750 0.771667 0.092500 0.136667
1 0.378750 0.900000 0.087500 0.166667
0 0.145000 0.695000 0.065000 0.116667
1 0.588750 0.801667 0.092500 0.143333
1 0.522500 0.716667 0.070000 0.140000
1 0.971250 0.608333 0.057500 0.156667
1 0.661250 0.453333 0.082500 0.146667
1 0.503750 0.408333 0.082500 0.130000
1 0.498750 0.245000 0.092500 0.123333
1 0.500000 0.116667 0.090000 0.086667
1 0.677500 0.240000 0.070000 0.146667
1 0.780000 0.240000 0.085000 0.133333
1 0.936250 0.305000 0.092500 0.143333
1 0.853750 0.166667 0.067500 0.120000
1 0.858750 0.071667 0.067500 0.090000
1 0.272500 0.475000 0.100000 0.136667
1 0.043750 0.325000 0.072500 0.103333
```

图 12 数据集标注示例

4.2 实验环境与模型训练

实验环境使用Ubuntu20.04 操作系统，使用GeForce GTX 1080Ti显卡进行运算，显存大小为11GB，CPU配置为Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz，CUDA版本为11.4.0，Pytorch版本为1.9.0，Python语言环境为3.7.4。

本实验总迭代次数为140 次，迭代批量大小设置为32，优化器选择SGD。模型训练时学习率使用Warmup[24] 训练预热，减缓模型在初始阶段对小批量数据的过拟合现象，避免模型振荡以便保证模型深层次的稳定性。在Warmup阶段，偏置层的学习率由0.1下降至0.01，其他的参数学习率由0增加至0.01，Warmup 结束之后，采用余弦退火学习算法[25] 对学习率进行更新。

4.3 评估指标与实验结果分析

本文评估指标采用平均精度(AP, Average Precision)、平均精度均值(mAP, mean Average Precision)以及每秒检测图片的帧数(FPS, Frames Per Second)这三种在目标检测算法中较为常见的评价指标来评估本文算法的性能。平均精度与精确率(Precision) 和召回率(Recall) 有关，精确率是指预测数据集中预测正确的正样本个数除以被模型预测为正样本的个数；召回率是指预测数据集中预测正确的正样本个数除以实际为正样本的个数。上述衡量指标的计算公式如下：

$$AP = \int_0^1 PdR \tag{14}$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N} \tag{15}$$

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{all\ \ detctions} \tag{16}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{all\ \ ground\ \ truths} \tag{17}$$

其中，AP值是指P-R曲线面积，本文采用插值计算的方法来计算公式 (14) 中的积分。公式 (15) 中mAP的值是通过所有类别的AP 求均值得到，N表示检测的类别总数，本实验中N=2，mAP的值越大，表示算法检测效果越好，识别精度越高。公式 (16) 和公式 (17) 中的TP、FP和FN分别表示正确检测框、误检框和漏检框的数量。

在训练140个迭代周期过程中，平均精度均值、精确率和召回率的变化曲线如图 13所示。从图 13中可以看出，模型在训练的过程中，在Warmup阶段

结束后的几个迭代周期中，平均精度均值、精确率和召回率有些许下降，随后随着余弦退火算法对学习率的调整，模型逐渐达到收敛状态。

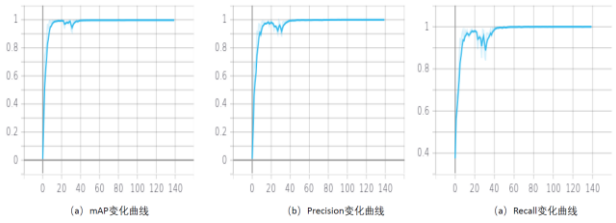


图 13 模型在数据集上的训练过程

为进一步验证本文算法的有效性，将本文算法与参考文献[23] 方法、AIZOO方法、原始YOLOv5算法在同一测试集上进行测试，各项性能指标比较结果如表 2所示。

表 2 不同算法性能对比结果

算法	AP/%		mAP/%	Times/s	FPS
	bad	good			
参考文献[23]	83.53	84.17	83.85	0.028	35.3
AIZOO	87.36	86.88	87.12	0.021	47.6
YOLOv5	89.49	91.16	90.33	0.024	41.6
本文算法	93.21	96.54	94.88	0.033	30.3

从表 2中可以看出，相较于参考文献[23] 方法、AIZOO方法和原始YOLOv5算法，本文算法在密集人群场景下对口罩这个小目标的检测表现效果更好，mAP值可以达到94.88%，在原始YOLOv5的基础上，bad和good类别的AP值分别提高了3.72%和5.38%，mAP值提高了4.55%。本文算法在检测速率上不及其他算法，FPS 为30.3，与原始YOLOv5 相比，FPS下降了11.3，检测单张图片的时间增加了9ms，由于实时检测一般要求检测帧率大于25帧/s，故本文算法仍能满足实时性要求。本文算法与参考文献[23] 方法、AIZOO 方法、原始YOLOv5 算法进行对比的检测效果如图 14所示。

图 14 检测效果对比图

（从上至下依次为：参考文献[23] 方法、AIZOO、原始YOLOv5、本文算法）

从图 14 中可以看出，参考文献[23] 方法在小目标异常角度、人脸区域有遮挡的条件下表现较差；AIZOO方法在检测效果上整体表现要稍好于参考文献[23] 方法，单帧检测时间最少，FPS 最高；原始YOLOv5 算法相较于前两种方法表现相对较好，但在一些小目标和存在遮挡条件下仍存在误判或者漏检的情况；与其他算法相比，本文算法在密集人群口罩佩戴检测效果中表现突出，检测精度有明显上升，在人群密集的条件下误判和漏检的效果明显下降，对小目标异常角度、人脸区域存在遮挡的鲁棒性明显提升。在图 14 (a1) 中，参考文献[23] 方法漏检了3个明显佩戴口罩的人脸目标，同时对4个佩戴口罩的小人脸目标做出了误判，图 14 (a2) 中AIZOO方法同样漏检了3个明显佩戴口罩的小人脸目标，图 14 (a3) 中原始YOLOv5算法对一个明显佩戴了口罩的小人脸目标做出了误判，图 14(a4) 中本文算法成功检测出了参考文献[23] 和AIZOO方法漏检的小人脸目标，同时成功检测出了明显未佩戴口罩的人脸小目标；在图 14 (b1) 中参考文献[23] 方法漏检了2个明显佩戴口罩的小人脸目标，同时对1个明显佩戴口罩的人脸目标误判，图 14 (b2) 中AIZOO方法漏检了3个未佩戴口罩的小人脸目标，图 14 (b3)中原始YOLOv5 算法漏检了1个未佩戴口罩的小人脸目标，同时对1个明显佩戴口罩的人脸目标

误判，图 14 (b4) 中本文算法成功检测出了11个明显佩戴口罩的小人脸目标和2个正确佩戴口罩的人脸目标；图 14 (c1) 中参考文献[23] 方法未检测出1个明显佩戴的人脸目标，图 14 (c2) 中AIZOO方法对明显佩戴口罩的1 个人脸目标做出了误判，图 14 (c3) 中原始YOLOv5 算法对2个明显佩戴口罩的人脸目标做出了误判，图 14 (c4) 本文算法成功检测到了为被参考文献[23] 方法检测到的正确佩戴口罩的人脸目标，并且对AIZOO方法和原始YOLOv5算法误判的明显佩戴口罩的人脸目标做出了正确判断; 图 14 (d1) 中参考文献[23] 方法对1个有遮挡的佩戴口罩的小目标人脸做出了误判，并且未检测出图片中间明显佩戴口罩的侧脸目标，图 14 (d2) 中AIZOO方法对1个被遮挡的佩戴口罩的小人脸目标做出了误判，图 14 (d3) 中原始YOLOv5算法未检测出被遮挡的佩戴口罩的小人脸目标，图 14 (d4) 中本文算法不仅成功检测出被遮挡的佩戴口罩的小人脸目标，而且对图片中间佩戴口罩的侧脸目标也做出了正确的判断。从图 14 (a4)(b4)(c4)(d4) 中可以看出，本文算法在人群密集且存在遮挡的场景下对于小目标的检测相较于其他算法检测效果明显，优势突出。

5 结束语

本文在原有YOLOv5算法的基础上，分别从Mosaic数据增强、特征提取器、损失函数和目标框回归四个方面进行改进，有效地增强了YOLOv5网络模型对小目标物体的检测精度，改进后的算法检测速率相较于原始YOLOv5算法有所降低，但仍能满足实时性的要求，可以直接应用在自动驾驶、医学图像、遥感图像分析和红外图像中的小目标检测等实际生活场景中。

参 考 文 献

[1] Najibi M, Samangouei P, Chellappa R, et al. Ssh: Single stage headless face detector[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4875-4884.

[2] Chen Q, Tang S, Yang Q, et al. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds[C]//2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2019: 514-524.

[3] Wang Z, Fu H, Wang L, et al. SCNet: Subdivision coding network for object detection based on 3D point cloud[J]. IEEE Access, 2019, 7: 120449-120462.

[4] Zhang L, Lin L, Liang X, et al. Is faster R-CNN doing well for pedestrian detection?[C]//European conference on computer vision. Springer, Cham, 2016: 443-457.

[5] Raghunandan A, Raghav P, Aradhya H V R. Object detection algorithms for video surveillance applications[C]//2018 International Conference on Communication and Signal Processing (ICCSP). IEEE, 2018: 0563-0568.

[6] 李红光,于若男,丁文锐.基于深度学习的小目标检测研究进展[J].航空学报,2021,42(07):107-125.

[7] Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. Springer, Cham, 2014: 740-755.

[8] Zhu Z, Liang D, Zhang S, et al. Traffic-sign detection and classification in the wild[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2110-2118.

[9] 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. 2014: 580-587.

[10] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[11] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[12] 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 and machine intelligence, 2016, 39(6): 1137-1149.

[13] 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.

[14] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector[C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[15] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.

[16] Redmon J, Farhadi A. Yolov3: An incremental

improvement [J]. arXiv preprint arXiv:1804.02767, 2018.

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

[18] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993-13000.

[19] 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. 2017: 2117-2125.

[20] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.

[21] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 658-666.

[22] Neubeck A, Van Gool L. Efficient non-maximum suppression[C]//18th International Conference on Pattern Recognition (ICPR'06). IEEE, 2006, 3: 850-855.

[23] 肖俊杰.基于YOLOv3和YCrCb的人脸口罩检测与规范佩戴识别[J].软件,2020,41(07):164-169.

[24] Xiong R, Yang Y, He D, et al. On layer normalization in the transformer architecture[C]. International Conference on Machine Learning, 2020: 10524-10533.

[25] Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts[J]. arXiv preprint arXiv:1608.03983, 2016.

编　辑