基于深度学习的口罩佩戴检测与跟踪

王　林，南改改
（西安理工大学 自动化与信息工程学院，陕西 西安 710048）

摘　要： 佩戴口罩可以有效预防病毒的传播，为减少通过人工方式检查口罩佩戴情况所消耗的大量人力资源，提出一种基于深度学习的口罩佩戴检测与跟踪方法，该方法分为检测和跟踪两个模块。检测模块在 YOLOv3 网络的基础上引入空间金字塔池化结构，实现不同尺度的特征融合；然后将损失函数改为 CIoU 损失，减少回归误差，提升检测精度，为后续跟踪模块提供良好的条件。跟踪模块采用多目标跟踪算法 Deep SORT，对检测到的目标进行实时跟踪，有效防止重复检测，改善被遮挡目标的跟踪效果。测试结果表明，该方法的检测速度为 38 f/s，平均精度值达到为 85.23%，相比原始 YOLOv3 算法提高了 4%，能达到实时检测口罩佩戴情况的效果。

关键词： 目标检测；目标跟踪；口罩佩戴检测；YOLOv3；Deep SORT
中图分类号： TP391.4　　**文献标识码：** A　　**DOI：** 10.16157/j.issn.0258-7998.211870

中文引用格式： 王林，南改改. 基于深度学习的口罩佩戴检测与跟踪[J].电子技术应用，2022，48(5)：21-26.
英文引用格式： Wang Lin，Nan Gaigai. Detection and tracking of mask wearing based on deep learning[J]. Application of Electronic Technique，2022，48(5)：21-26.

Detection and tracking of mask wearing based on deep learning

Wang Lin，Nan Gaigai
(School of Automation and Information Engineering，Xi′an University of Technology，Xi′an 710048，China)

Abstract： Wearing a mask can effectively prevent the spread of the virus. In order to reduce the consumption of a large number of human resources in manual inspection of mask wearing, this paper proposes a method of mask wearing detection and tracking based on deep learning, which is divided into two modules: detection and tracking. Based on the YOLOv3 network, the spatial pyramid pooling structure is introduced into the detection module to realize the feature fusion at different scales, then the loss function is changed to CIoU loss to reduce the regression error improve detection accuracy, and provides good conditions for the subsequent tracking module. The tracking module adopts the multiple object tracking algorithm Deep SORT to track the detected objects in actual time, which can effectively avoid repeated detection and better the tracking effect of the occluded targets. The test results indicate that the detection velocity of this way is 38 f/s, and the average accuracy value is 85.23%, which is 4% higher than the original YOLOV3 algorithm, and can achieve the effect of real-time detection of mask wearing.
Key words： object detection；object tracking；mask wearing detection；YOLOv3；Deep SORT

0 引言

近年来，人们所生活的环境空气污染程度日益严重，由此引发了一系列的疾病，尤其严重的是尘肺职业病，发病率在持续增长。再加上，2019 年 12 月，新型冠状病毒肺炎疫情(COVID-19)[1]的爆发，使得人们不得不对此采取一定的措施。大量证据表明，佩戴口罩能有效缓解类似疾病的发生和传播，所以在医院、后厨和化工厂等特殊场所佩戴口罩已是必然。然而，佩戴口罩不仅需要个人自觉遵守，更需要采取一定的措施进行监督管理。

目标检测和目标跟踪[2]是计算机视觉中的一个基础内容。基于深度学习的目标检测方法大致可以分为两大类：一类是两阶段检测器，最典型的是 R-CNN 系列，包括 R-CNN[3]、Fast R-CNN[4]和 Faster R-CNN[5]；另一类是单阶段检测器，包括 YOLO[6]、SSD[7]和 RetinaNet[8]等。其中，YOLOv3[9]是 YOLO 系列中最为广泛使用的目标检测算法，具有较好的识别速度和检测精度。目标跟踪主要分为生成式方法和判别式方法。前者主要是对目标进行建模，具有代表性的算法有 CSK[10]和 IVT[11]等；判别式方法相比生成式方法最大的区别在于其不仅学习目标本身，更关注如何将前景和背景分开，具有代表性的算法有 KCF[12]和 TLD[13]等。作为多目标跟踪的 Deep SORT[14]算法是在 SORT 目标跟踪算法基础上的改进，使用了逐帧数据关联和递归卡尔曼滤波的传统单假设跟踪方法，在实时的目标跟踪过程中具有较好的性能。

在实际的口罩佩戴检测过程中，通过人工方式检查口罩佩戴情况会消耗大量人力资源，而且在人群密集的

区域，一般会存在遮挡和小目标检测困难的问题，从而导致漏检。对于视频数据，若对每一帧图像都进行人脸口罩佩戴检测，不仅浪费时间，而且会耗费大量的计算资源，难以满足实时性的要求。目前所提出的专门用于口罩佩戴检测的算法较少，而且这些算法只是检测，并没有涉及跟踪。

基于此本文提出一种基于深度学习的方法，对特定场所内的人员口罩佩戴情况进行实时检测和跟踪，该方法分为检测和跟踪两个模块。由于 YOLOv3 算法对小目标和遮挡物体的检测效果并不理想，因此检测模块本文尝试在 YOLOv3 的基础上引入空间金字塔池化结构，然后将损失函数由 IoU 改为 CIoU，来改善遮挡和小目标的问题，并在此基础上提升检测精度。跟踪模块采用多目标跟踪算法 Deep SORT，与改进的 YOLOv3 算法结合，先对视频第一帧中的人脸目标进行检测定位，然后再进行跟踪，确定后续帧中的运动信息，从而有效防止重复检测，改善被遮挡目标的跟踪效果，提升检测速度。

1 目标检测

目标检测的工作是在提供的输入图像上，通过计算机找出是否有规定类别的目标，如果计算机认为有则进一步给出该目标的位置和大小，这个位置和大小通常用矩形边框左上角的坐标及长和宽来表示。

1.1 YOLOv3 算法模型

YOLOv3 是以 YOLOv1 和 YOLOv2 为基础来进行改进的，不仅保证了速度的优势，而且提升了检测精度，同时增强了对小目标的检测能力。YOLOv3 采用含有 53 个卷积层的 Darknet-53，所达到的效果比 ResNet-152 网络更好。YOLOv3 的网络结构如图 1 所示。

从图 1 中可以看出，与 ResNet 不同的是 Darknet-53 网络没有最大池化层，该网络中所有的下采样基本上都是通过卷积层来实现的，从而提升 YOLOv3 的检测效果。由于 Darknet-53 网络所采用的卷积核个数少一些，因此参数就少一些，使得运算量也减少，以此提高了检测速度。其中的 Convolutional 不单是一个卷积层，而是一个普通的卷积加上归一化层(BN Layer)，再加上激活函数层(Leaky ReLU Layer)组成的，Convolutional Set 是由 5 个卷积核不同的 Convolutional 组成的。残差块(Residual)是在 2 个 Convolutional 的基础上加了一个拼接(Add)操作。

1.2 改进 YOLOv3 算法

针对口罩佩戴检测存在的问题，本文以 YOLOv3 算法为基础引入空间金字塔池化(Spatial Pyramid Pooling，SPP)网络，并对损失函数进行改进，从而提升后续跟踪算法的性能。

1.2.1 引入空间金字塔池化

SPP-net[15]提出了一种叫做空间金字塔池化的方式，可以允许任意大小的图像输入网络，从而对精度有一个更好的保证，保留了更多的输入信息，同时对计算速度起到一个极大的加快作用。本文在 DarkNet-53 网络第一个预测特征层的 Conventional 中引入 SPP 结构，从而获取更丰富的局部特征，引入的空间金字塔池化网络结构如图 2 所示。

由图 2 可以看出其具体步骤为：先对 13×13×1 024 的输入特征图进行 3 次 Conventional 操作；然后分别利用 13×13、9×9 和 5×5 的池化核对池化层(Pooling)进行最大池化下采样，步距(stride)都为 1，意味着在池化之前要对

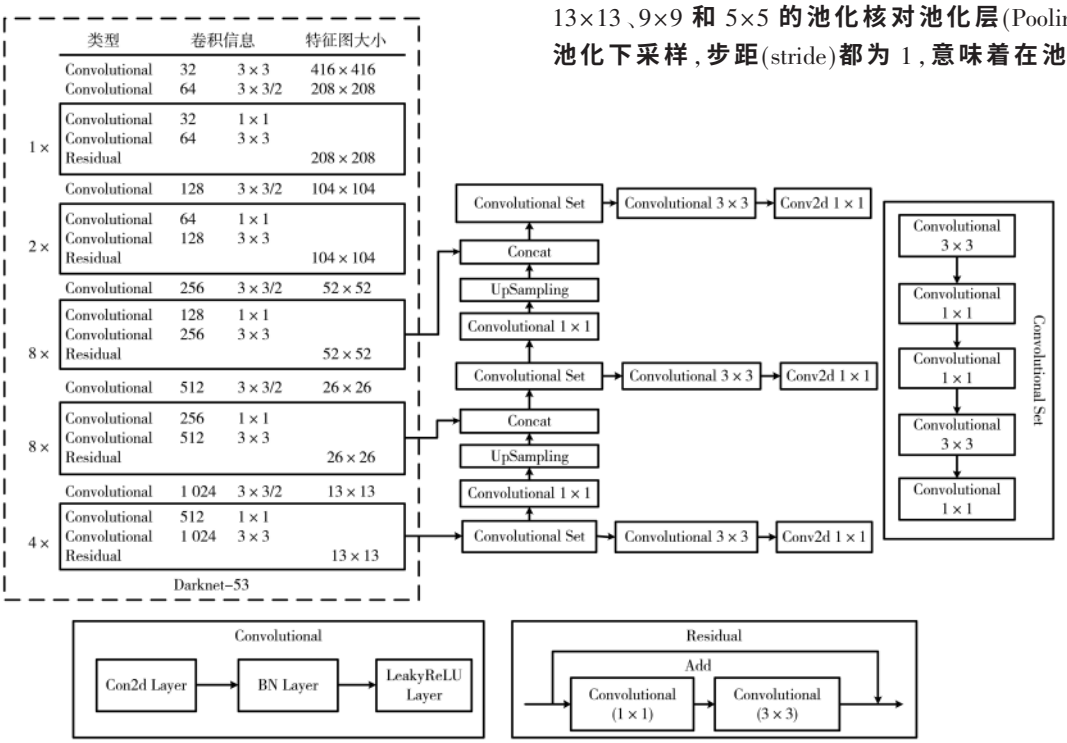


图 1 YOLOv3 网络结构

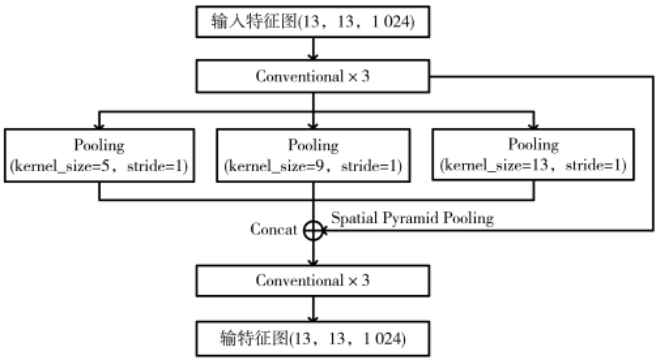


图 2　改进的空间金字塔池化网络结构

特征矩阵进行填充(padding)，使最大下采样之后得到的特征图的高度和宽度不变；最后将输入的特征图和池化处理后得到的特征图进行拼接(Concat)，然后再经过 3 次 Conventional 操作，从而得到输出特征图。通过 SPP 实现了不同尺度的特征融合。

1.2.2 改进损失函数

YOLOv3 的损失函数由三部分组成，分别是置信度损失 $L_{conf}(o,c)$、分类损失 $L_{cla}(O,C)$ 以及定位损失 $L_{loc}(\boldsymbol{l},\boldsymbol{g})$，其计算公式如式(1)所示：

$$L(o,c,O,C,\boldsymbol{l},\boldsymbol{g})=\lambda_1 L_{conf}(o,c)+\lambda_2 L_{cla}(O,C)+\lambda_3 L_{loc}(\boldsymbol{l},\boldsymbol{g}) \tag{1}$$

式中，λ_1、λ_2、λ_3 为平衡系数；o 表示的是预测目标边界框与真实目标边界框的交并比；O 表示预测目标边界框中是否存在目标，只有 0 和 1 两个值；c 和 C 为预测值；$\boldsymbol{l}$ 表示预测矩形框的坐标偏移量，g 表示与之匹配的真实框与默认框之间的坐标偏移量。

其中置信度损失和类别损失使用的是二值交叉熵损失(Binary Cross Entropy)，定位损失使用差值平方和的计算方法。由于定位损失不能很好地反映两个目标边界框的重合程度，因此引入交并比损失(IoU loss)，其计算公式如式(2)所示：

$$L_{IoU}=1-\mathrm{IoU} \tag{2}$$

$$\mathrm{IoU}=\frac{|A\cap B|}{|A\cup B|} \tag{3}$$

式中，A 和 B 分别表示预测框和真实框。IoU loss 能够很好地反映出两个矩形框的重合程度，而且重合程度与矩形框的尺度无关，但是当两个目标边界框不相交时 IoU 为 0，就无法传播损失。而 CIoU 同时考虑到目标边界框的重叠部分的面积、中心点距离以及长宽比的影响，其计算公式如下：

$$\mathrm{CIoU}=\mathrm{IoU}-\left(\frac{\rho^2(\boldsymbol{b},\boldsymbol{b}^{gt})}{C^2}+\alpha v\right) \tag{4}$$

$$v=\frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}}-\arctan\frac{w}{h}\right)^2 \tag{5}$$

$$\alpha=\frac{v}{(1-\mathrm{IoU})+v} \tag{6}$$

其中，b 是预测目标边界框的中心坐标，b^{gt} 是真实目标边界框中心点的坐标，$\rho^2(b,b^{gt})$ 表示 b 和 b^{gt} 之间欧氏距离的平方，C 是两个目标边界框最小外接矩形的对角线长度，αv 表示的目标边界框的长宽比信息。相应的 CIoU Loss 的计算公式如下：

$$L_{CIoU}=1-\mathrm{CIoU} \tag{7}$$

使用 CIoU Loss 可以很好地提升目标检测的准确性。

2 目标跟踪

跟踪的主要思路就是检测加跟踪，将检测的结果(bounding box、confidence 和 feature)作为输入，其中 confidence 主要是用来筛选检测框，将 bounding box 和 feature 与跟踪器进行匹配计算。本文采用多目标跟踪算法 Deep SORT 来对检测到的目标进行跟踪，有效防止重复检测，改善被遮挡目标的追踪效果，从而提高检测精度和速度。

Deep SORT 对视频流中的每一帧进行检测，通过检测出的外形和运动特征来跟踪目标。Deep SORT 在描述运动状态时使用一个 8 维的向量 $(\mu,v,\gamma,h,\dot{x},\dot{y},\dot{\gamma},\dot{h})$ 来进行轨迹的刻画，其中 (μ,v) 表示边界框的中心位置的坐标，h 为高度，γ 为横纵比，剩余变量 $(\dot{x},\dot{y},\dot{\gamma},\dot{h})$ 则为图像坐标系中的速度信息。轨迹的预测更新使用的是采用线性预测模型和匀速模型的卡尔曼(Kalman)滤波器，预测结果为 (μ,v,γ,h)。Deep SORT 采用匈牙利算法将预测后的轨迹和当前帧中的目标进行匹配，采用的匹配方法是级联匹配和 IoU 匹配，从而保证跟踪的持续性。

运动信息的匹配程度采用检测框和跟踪在卡尔曼滤波器的预测框之间的马氏距离来刻画，其表达式如下：

$$d^{(1)}(i,j)=(\boldsymbol{d}_j-\boldsymbol{y}_i)^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{d}_j-\boldsymbol{y}_i) \tag{8}$$

其中，$d^{(1)}(i,j)$ 表示第 j 个检测框和第 i 条轨迹之间的运动关联程度，其中 $\boldsymbol{d}_j$ 和 $\boldsymbol{y}_i$ 分别表示第 j 个检测框位置和第 i 个追踪器预测目标的位置，$\boldsymbol{S}_i$ 表示他们之间的协方差矩阵。

Deep SORT 添加了一个深度学习的特征提取网络来匹配更新存放特征图的列表，计算每一帧追踪器成功匹配的特征集与目标检测框的特征向量之间的最小余弦距离，其计算公式如下：

$$d^{(2)}(i,j)=\min\{1-\boldsymbol{r}_j^{\mathrm{T}}\boldsymbol{r}_k^{(i)}\,|\,\boldsymbol{r}_k^{(i)}\in\boldsymbol{R}_i\} \tag{9}$$

其中，$d^{(2)}(i,j)$ 表示第 i 个追踪器匹配的特征集与第 j 个检测框之间的最小余弦距离，$\boldsymbol{r}_j$ 和 $\boldsymbol{R}_i$ 分别表示其所对应的特征向量和特征向量集。最后将以上两种匹配方式的线性加权和作为最终的匹配度，其表达式如下：

$$c_{ij}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j) \tag{10}$$

其中，λ 表示权重系数。

3 实验结果与分析

3.1 网络训练

实验硬件环境：采用 4 块 NVIDIA 1080 Ti 的显卡，

128 GB 内存，处理器采用两颗 Intel Xeon E5-2640v4 的处理器。软件环境是在 CentOS7.0 的操作系统下，采用 Python 计算机语言，深度学习框架为 Pytorch。

本文按照 1:9 的比例划分测试集和训练集的数目，采用 Adam 优化器对网络进行优化，动量参数为 0.92。训练网络模型过程采用迁移学习的思想，分为冻结训练和解冻训练。先冻结一部分进行训练，学习率设为 0.001，batch_size 为 8，训练 110 个轮次(Epoch)；然后再进行解冻训练，学习率设为 0.000 1，batch_size 为 4，共训练 220 个轮次(Epoch)。其中损失值收敛曲线如图 3 所示。

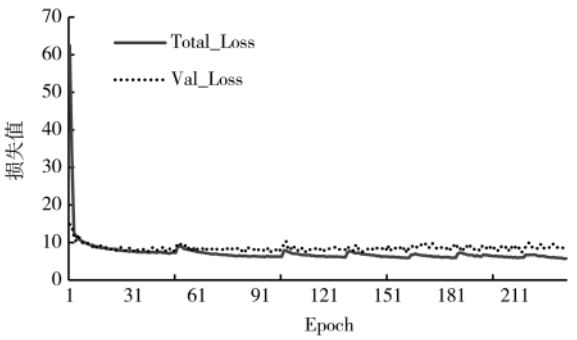


图 3 损失曲线

从图 3 可以看出，训练初期总损失(Total_Loss)和验证损失(Val_Loss)都下降较快，随着训练次数的增加逐渐趋于稳定，最后的损失值为 5.6 左右。另外，在同等的实验条件下训练 YOLOv3 和 SSD 的网络模型，保证训练方式不变，从而进行更好的分析比较。

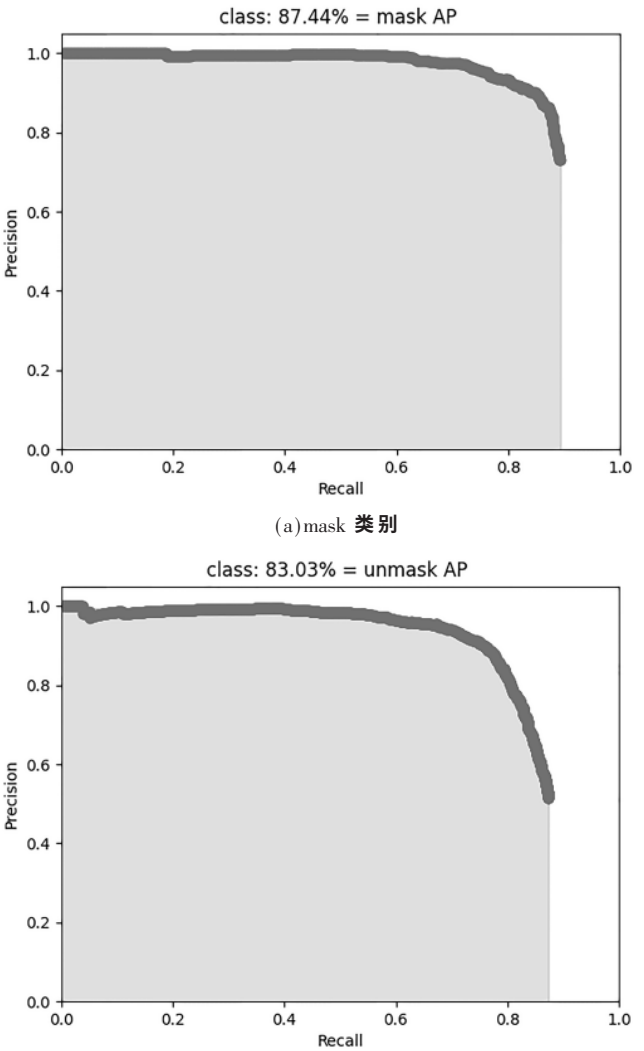
3.2 数据集

本文使用部分开源的人脸图片数据集 MAFA(Mask Faces)[16]和 Wilder Face，并通过网络爬虫和视频拍摄等自制了佩戴口罩的人脸数据集，然后删除部分质量较差的数据，最后通过数据增强的方式将数据集扩展到9 240 张，其中包括 5 000 多张佩戴口罩的数据。数据集包括在不同场景下人脸佩戴口罩和未佩戴口罩两种情况，将数据集格式转换成 VOC2007 数据集所要求的格式，并用标注工具 LableImg 进行数据标注，分为 mask 和 unmask 两种标注类别，数据集标注情况如图 4 所示。

图 4 数据集标注示例

3.3 结果分析

本文引入空间金字塔池化结构，获得丰富的局部特征信息，并采用 CIoU 作为目标边界框损失函数，提高了收敛速度和回归精度，且降低了预测误差。最后根据召回率和精确度的值绘制 *P*-*R* 曲线，如图 5 所示。横纵坐标分别表示召回率(Recall)和精确度(Precision)，平均精确度 AP 表示为曲线下的面积大小。mAP 是平均精度均值，即对所有类别的 AP 取平均值，反映的是网络模型的整体性能，是模型性能评估的重要指标。



(a)mask 类别

(b)unmask 类别

图 5 各类目标 *P*-*R* 曲线

由图 5 可以看出，多目标检测的平均精度达到 85.23%，佩戴口罩的类别达到 87.44%，与原始 YOLOv3 模型相比均有提高，说明检测效果有所提升。

为了验证口罩佩戴检测模型的性能，在相同的实验环境下用同样的数据集对 SSD 和 YOLOv3 原网络进行了训练。在得到对应的检测模型上再用相同的数据进行测试，结果如表 1 所示。

表 1　各模型检测性能指标对比

模型	AP/%		mAP/%	FPS/(f/s)
	mask	unmask		
SSD	81.64	75.20	78.42	35
YOLOv3	84.52	78.02	81.27	32
本文算法	87.44	83.03	85.23	38

本文以平均检测精度(mAP)和帧率(FPS)作为评价指标。由表 1 可以看出，本文所提出的算法要比 SSD 和 YOLOv3 算法的检测精度高，且检测速度也表现较好。

针对侧脸、遮挡和密集人群 3 种情况选取部分数据，分别利用 SSD、YOLOv3 和本文所提出的检测模型对其进行测试，对比结果如图 6 所示。

由图 6 可以看出，SSD 算法对小目标的检测效果并不好；YOLOv3 算法相比 SSD 有所提升，但是容易漏检，有些很明显的目标反而没有被检测到；本文提出的算法对小目标和遮挡都表现出良好的效果，更加适合用来进行口罩佩戴检测，为后续的跟踪模块提供了不错的条件。

口罩佩戴检测与跟踪算法的主要流程为：首先将视频流输入，运行检测模型得到检测框；然后通过跟踪模型计算最小余弦向量及马氏距离并融合为关联矩阵；最后进行匹配跟踪。图 7 是对视频数据进行测试结果的截图展示。

图 7 中白色框是标有置信度的检测框，灰色框是标

图 7　视频人脸口罩佩戴检测与跟踪结果

(a)SSD 算法　(b)YOLOv3 算法　(c)本文算法

图 6　不同算法检测结果对比

《电子技术应用》http://www.chinaaet.com

有不同 ID 号的跟踪框，并且对佩戴口罩的人员绘制跟踪轨迹。可以看出，本文提出的算法无论在检测还是跟踪模型上都取得了不错的效果。

4 结论

本文提出了一种基于深度学习的口罩佩戴检测与跟踪方法。检测模块是在 YOLOv3 网络的基础上，引入空间金字塔池化结构，实现了不同尺度的特征融合，提高了小目标的检测准确率；然后将损失函数由 IoU 改为 CIoU，减少回归误差，进一步提升了检测性能，同时也有助于提升后续跟踪的效果。跟踪模块采用多目标跟踪算法 Deep SORT 与检测模块相结合，来对检测到的目标进行跟踪，有效避免了重复检测，同时也能解决部分遮挡问题。测试结果表明，本文提出的算法可以有效提高检测的精度和速度，平均精度值达到 85.23%，检测速度达到 38 f/s。针对该算法对视频流进行测试，发现其在检测与跟踪方面都取得了不错的效果，对传播性疾病的防控工作具有很好的应用前景。在实用性方面可以考虑加入报警模块，对未佩戴口罩的行为及时发出警报，以供后台人员进行处理。

参考文献

[1] 赵文明，宋述慧，陈梅丽，等.2019 新型冠状病毒信息库[J].遗传，2020，42(2)：212-221.

[2] 王福建，张俊，卢国权，等.基于 YOLO 的车辆信息检测和跟踪系统[J].工业控制计算机，2018，31(7)：89-91.

[3] GIRSHICK R，DONAHUE J，DARRELL T，et al.Region-based convolutional networks for accurate object detection and segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2016，38(1)：142-158.

[4] GIRSHICK R.Fast R-CNN[C]//Proceedings of 2015 IEEE International Conference on Computer Vision and Pattern Recognition.Washington D.C.，USA：IEEE Press，2015：1440-1448.

[5] REN S Q，HE K M，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.

[6] 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.Washington D.C.，USA：IEEE Press，2016：779-788.

[7] LIU W，ANGUELOV D，ERHAN D，et al.SSD：single shot multi-box detector[C]//Proceedings of 2016 European Conference on Computer Vision.Berlin，Germany：Springer，2016：21-37.

[8] LIN T Y，GOYAL P，GIRSHICK R，et al.Focal loss for dense object detection[J].IEEE Transactions on Pattern Analysis & Machine Intelligence，2017，PP(99)：2999-3007.

[9] REDMON J，FARHADI A.YOLOV3：an incremental improvement[C]//Proceedings of 2018 IEEE Conference on Computer Vision and Pattern Recognition.Washington D.C.，USA：IEEE Press，2018：1-6.

[10] HENRIQUES J F，CASEIRO R，MARTINS P，et al.Exploiting the circulant structure of tracking-by-detection with kernels[C]//ECCV，2012.

[11] ROSS D A，LIM J，LIN R S，et al.Incremental learning for robust visual tracking[J].International Journal of Computer Vision，2008，77(1-3)：125-141.

[12] HENRIQUES J F，CASEIRO R，MARTINS P，et al.High-speed tracking with kernelized correlation filters[J].IEEE Transactions on Pattern Analysis and Machine Intelligence，2015，37(3)：583-596.

[13] WOJKE N，BEWLEY A，PAULUS D.Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing(ICIP)，2017：3645-3649.

[14] KALAL Z，MIKOLAJCZYK K，MATAS J.Tracking-learning-detection[J].Machine Intelligence，2012，34(7)：1409-1422.

[15] He Kaiming，Zhang Xiangyu，Ren Shaoqing，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.

[16] GE S，LI J，YE Q，et al.Detecting masked faces in the wild with LLE-CNNs[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition，2017：426-434.

（收稿日期：2021-06-18）

作者简介：

王林(1962-)，男，博士，教授，主要研究方向：深度学习、计算机视觉。

南改改(1995-)，女，硕士研究生，主要研究方向：深度学习、计算机视觉。

扫码下载电子文档

版权声明