第 59 卷 第 8 期/2022 年 4 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于改进YOLOv4的行人鞋部检测算法

杨智雄，唐云祁*，张家钧，耿鹏志

中国人民公安大学侦查学院，北京 100038

摘要 结合现场鞋印和周边监控视频锁定犯罪嫌疑人是公安机关刑事侦查破案的一种重要技战法，然而该技战法自动化程度低、耗时耗力，限制了应用范围。针对这一问题，本文提出一种基于YOLOv4算法的目标检测方法，可实现对监控视频下行人鞋部的自动检测。根据行人鞋部区域的特点，首先使用K-means聚类算法获得先验框尺度，并确定其数量；然后根据构建的数据集选取合适的检测层以强化对鞋部特征的学习；最后，通过多尺度特征融合，将调整后的空间金字塔池化结构迁移到剪枝后的网络内，增强模型的学习能力。结果表明，提出的YOLOv4_shoe算法训练权重仅为39.56 MB，参数量约为原模型的六分之一，mAP值达到了97.93%，比原YOLOv4模型提升了2.07%。

关键词 图像处理；鞋部检测；YOLOv4；特征融合；空间金字塔池化；视频监控

中图分类号 TP391.4 **文献标志码** A **doi**：10.3788/LOP202259.0810007

Detection Algorithm of Pedestrian Shoe Area Based on Improved YOLOv4

Yang Zhixiong, Tang Yunqi*, Zhang Jiajun, Geng Pengzhi

School of Investigation, People's Public Security University of China, Beijing 100038, *China*

Abstract One of the important tactics used by the public security bureau in a criminal investigation is to combine the related surveillance video and shoeprints on the spot to identify criminal suspects. However, the low-level automation of such a method is so labor-intensive and time-consuming, which limits its application. Therefore, this paper proposes an object detection method based on the YOLOv4 algorithm to realize the automatic detection of pedestrian shoes in surveillance video. According to the characteristics of the pedestrian shoe area, first, the K-means clustering algorithm is used to determine the scale of the anchor box and confirm its quantity; second, an appropriate detection layer was selected based on the datasets in this paper to improve the learning of shoe features; finally, a multifeature fusion method is used and the adjusted spatial pyramid pooling structure is transferred into the pruned network to improve the learning ability of the model. The experimental results demonstrate that the training weight of the YOLOv4_shoe algorithm proposed is only 39.56 MB, which is approximately one-sixth of the original model; and its mean average precision reaches 97.93%, which is 2.07% higher than that of the original YOLOv4 model.

Key words image processing; shoes detection; YOLOv4; feature fusion; spatial pyramid pooling; video surveillance

收稿日期：2021-03-22；**修回日期**：2021-04-17；**录用日期**：2021-04-28

基金项目：公安部技术研究计划项目(2020JSYJC21)、中央高校基本科研业务费项目(2021JKF203)

通信作者：*tangyunqi@ppsuc.edu.cn

1 引 言

随着我国经济与科技水平的提升，视频监控技术在公共场合广泛应用，人们的日常行为很容易被视频监控捕捉并记录下来。在公安领域，视频监控作为一种防控措施，具有准确性、客观性与关联性等特点，对监控视频内的信息进行深度挖掘，往往能给侦查破案提供线索，指明方向。例如“监控＋鞋印”技战法，根据作案人遗留在现场的鞋印痕迹，检索对比确定嫌疑人作案时所穿鞋型，然后通过筛查周边的监控视频实现鞋印与监控影像的关联。这种技战法由袁楚平等[1]提出，在其他侦查方案受阻后这种技战法提供了新的侦查思路，并于实战中取得了积极的效果。如2015年4月，广西某县发生一起三轮摩托车女司机被杀案，技术人员以摩托车后座垫板上的两枚残缺鞋印为着手点，通过筛查视频与大量的走访排查，成功从监控视频中确定了嫌疑人的身份[2]。虽然这种技战法颇有实战价值，但由于其依赖于人工筛查视频，耗时耗力，因此无法成为侦查破案的首选方案[3]。为了提高该技战法的自动化程度，更好发挥其实战价值，首先就是要实现监控下自动检测行人鞋部，这可为后续分类检索鞋型、锁定嫌疑人等工作奠定基础。

随着深度学习的发展，目标检测技术越来越成熟，但是目前关于监控下对行人鞋部检测的问题研究较少。耿鹏志等[4]提出一种基于SSD模型的鞋部检测算法，但是由于存在通过人工修改图像尺寸与先验框的问题，且其制作的实验数据库缺乏复杂背景、部分遮挡、不同行走方向等干扰因素下的数据，具有一定的局限性。此外，监控视角下行人鞋部属于小目标，导致检测结果分辨率低、图像模糊、有效信息少，这也为行人鞋部的检测工作增加了难度。

本文以目标检测算法YOLOv4为基础，提出面向行人鞋部的YOLOv4_shoe检测模型，实现对监控视频中行人鞋部的快速自动检测。首先在实验室模拟犯罪现场周边环境，收集实验所需的视频数据，然后对视频进行格式转换、图像分帧、标签分类，构建了包含3062张图像的实验数据库。本文提出的YOLOv4_shoe检测模型与原YOLOv4相比，设置了匹配数据集的先验框尺度，减少了先验框数量；并通过实验选取了合适的检测层，删去16，32倍下采样层的检测头以精简网络；使用多尺度特征融合以及调整后的空间金字塔池化结构，增强了模型的检测性能。

2 本文方法

2.1 目标检测

目标检测是计算机视觉领域的一项基本任务，旨在解决目标的定位与分类问题，其目的是获得图像中感兴趣目标的边界框位置信息与类别信息。目标检测包括基于传统手工特征的目标检测方法与基于深度学习的目标检测方法[5]。

在早期计算资源有限的环境下，目标检测算法大多是通过手工提取特征方法构建。由于缺乏对图像信息进行结构化处理的有效途径，往往通过复杂的特征表示方法以及各种加速技术来尽可能充分利用计算资源，代表性的检测模型有Viola-Jones(VJ)[6-7]检测器、方向梯度直方图(HOG)特征描述器[8]、基于可变形部件的目标检测模型(DPM)[9]等等。传统的目标检测方法由于滑动窗口与手工特征的局限性，逐步暴露出其迁移性差、泛化性弱等缺点。

随着深度学习技术的发展，目标检测技术取得了显著的进步，检测效果大大提升。基于深度学习的目标检测方法主要有两种，一是Two-stage方法，即把检测分为定位与分类两个步骤，代表算法有R-CNN[10]，Fast R-CNN[11]，Faster R-CNN[12]，Mask R-CNN[13]等，其思路是首先生成候选框区域，然后运用卷积神经网络，对候选框区域的特征向量进行分类回归。由于需要对每一个可能包含检测对象的区域进行检测，因此此方法准确度高，但是冗余的计算量也限制了检测速度，不符合实时检测的要求；另一种为One-stage方法，不需要生成候选区域，直接通过回归即可完成定位与分类任务，代表算法有YOLO[14-17]，SSD[18]，Retina-Net[19]等，此方法检测准确度总体较Two-stage方法低，但是检测速度快、效率高。

根据公安侦查破案实战中对效率的要求，本文将YOLOv4算法引入视频监控下行人鞋部的检测工作中，不仅可以帮助侦查人员更高效地开展检测工作，也可以避免人工筛查的主观因素造成的干扰，从而加快侦查破案的进程。

2.2 YOLOv4简介

YOLOv4算法是在YOLO系列算法基础上，通过大量实验与使用深度学习的小技巧，增强了算法的检测效果。其检测思路是将输入的图像分为76×76，38×38，19×19的网格，分别对应网络的

8倍下采样层、16倍下采样层和32倍下采样层，用于检测不同尺度的目标。网络共有9种大小不同的先验框，不同检测层下的网格分别对应3个先验框，模型最终通过多尺度预测获得较为综合的检测性能。

原始YOLOv4模型的网络结构如图1所示。相比于其他检测网络，YOLOv4的backbone部分使用了CSPDarknet53，具有较强的学习能力，使得模型在轻量化的同时保持准确性；在neck部分主要采用了Spatial Pyramid Pooling(SPP)[20]以及Feature Pyramid Networks(FPN)[21]+Path Aggregation Network(PAN)[22]结构来更好地提取融合特征；YOLO-head部分采用了CIoU_Loss与DIOU_nms算法，使检测框回归的速度与精度得到了提升[23]。

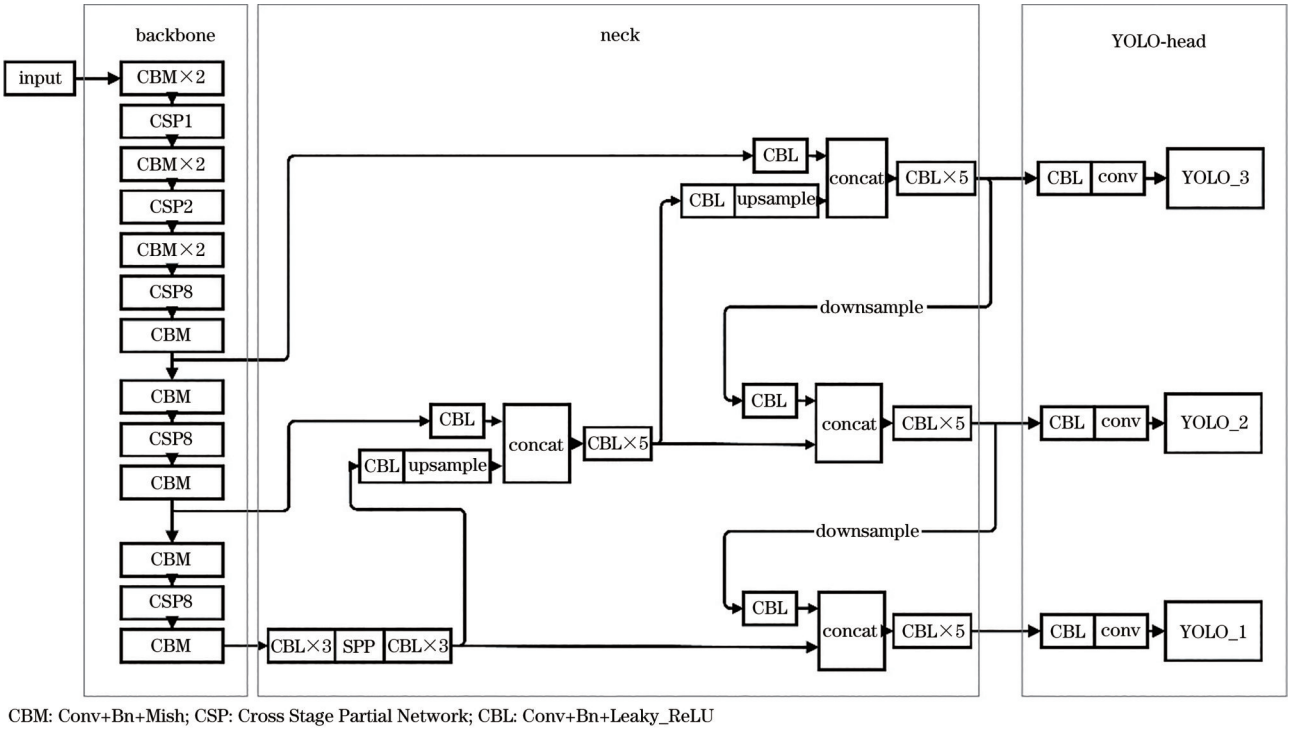

图1　YOLOv4网络结构

Fig. 1　YOLOv4 network structure

YOLOv4模型的损失函数是在YOLOv3的基础上进一步修改得到的，用CIoU误差代替了均方误差作为预测框的预测误差。模型总损失函数为$L=L_{\mathrm{loc}}+L_{\mathrm{obj}}+L_{\mathrm{cls}}$。其中$L_{\mathrm{loc}}$、$L_{\mathrm{obj}}$、$L_{\mathrm{cls}}$分别表示物体位置损失函数、置信度损失函数与分类损失函数[24]。

2.3　面向鞋部区域检测的YOLOv4算法改进

2.3.1　先验框尺寸与数量的设置

YOLOv4模型内的先验框是预先在公开数据集上聚类得到的，其面向的目标大小形状不一，差异较大，导致先验框的尺寸也各不相同，因此无法在训练中直接使用默认的先验框。为了更好地匹配检测对象，使用K-means聚类算法获取鞋部标签数据集，生成更符合鞋部的先验框尺寸。传统的K-means聚类算法通过欧氏距离函数进行聚类，但是在生成先验框过程中，较大的先验框会产生更显著的误差，不利于获取准确并有代表性的先验框尺度。因此为避免尺寸差异带来的影响，使用聚类中心与标签中真实框的IoU值作为聚类相似度的评判标准，这样可以保证获得的先验框与鞋部的形状、大小更加匹配。聚类距离函数计算方法为$d(m,n)=1-\mathrm{IoU}(m,n)$，其中，$m$，$n$分别代表边界框尺度、聚类中心。

值得注意的是，先验框数量会影响模型的性能，选取较多的先验框有利于更好地匹配目标，从而提高模型的召回率，但是先验框数量过多时也会导致计算量增加，从而影响模型的检测效率，因此实验前要根据检测对象的具体情况选取数量适当的先验框。

2.3.2　检测层的选取

在目标检测的过程中，主干特征提取网络对检测的性能影响很大，随着神经网络深度的变化，能够提取到不同层次的图像特征[25]。网络较浅时，有

利于获取图像的位置、纹理、边缘等细节特征；而网络较深时，则有利于获取图像的语义信息。YOLOv4算法在主干网络较浅时，会在图像上生成更细密的网格，有利于检测小目标；当主干网络较深时，图像网格则会更加粗略，不利于小目标检测。因此，本文结合原YOLOv4模型主干网络的特点，通过选取不同下采样层作为检测层，对比了不同检测层对检测性能的影响，最后通过实验选取了合适的检测层，保证网络能更好地学习小目标的特征。

2.3.3　多尺度特征融合

原YOLOv4模型的neck部分，运用FPN+PAN结构提取不同尺度的融合特征，通过自顶向下与自底向上两个方向的特征融合，加强了语义特征与定位特征的传输，从而增强了三个尺度检测层的性能。而本文针对检测目标较小的情况，通过简化多尺度特征融合的方式来提高模型的检测性能[26]。改进后的模型只使用一个检测层用于行人鞋部的检测工作，同时保留了更深的下采样层的网络分支，在模型的neck部分通过上采样将网络提取到的多层次特征进行融合，保证网络既可以学习到包含细节与位置信息的低层特征图，也可以融合包含语义信息的高层特征图。

2.3.4　空间金字塔池化结构

原YOLOv4模型在neck部分使用了空间金字塔池化结构(SPP)，其结构如图2所示。在特征图进入SPP结构后，分别由1×1，5×5，9×9和13×13四个尺度的内核进行最大池化操作，与原来单纯运用一个尺度的$n\times n$最大池化的方式相比，能够更有效地丰富特征图的表达能力，最终获取到特征图局部与全局的感受野信息。改进后的模型继续引用了SPP结构，但是，针对实验检测对象的特殊性，如果内核池化的尺度过大，会导致体积较小的鞋部信息在池化过程中丢失，因此，实验尝试对SPP结构进行适配性的调整，以提高模型对数据集的检测能力。

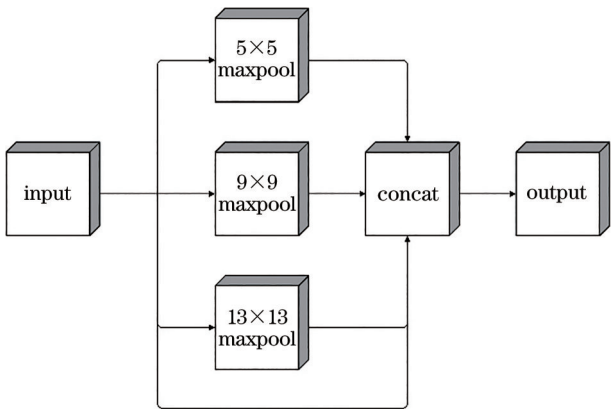

图2　空间金字塔池化结构

Fig. 2　Structure of spatial pyramid pooling

3　实验及结果分析

3.1　行人鞋部数据集的构建

实验采用的数据集来自足迹实验室，由13名志愿者(男性9名，女性4名)的视频数据制作而成，包括13双不同种类的鞋型。实验室一共使用了5种不同颜色、花纹的背景来模拟各类路面情况，以增强实验样本的丰富性。采集流程为：志愿者依次在铺设的背景纸上自然行走，沿背景纸铺设的路线从A点出发行进至B点再返回，如此重复两次。同时通过与志愿者行走方向呈0°、90°、180°三个角度的监控摄像头实时录制，每个人的视频时长大约为30 s，实验采集环境及行走路线如图3所示。

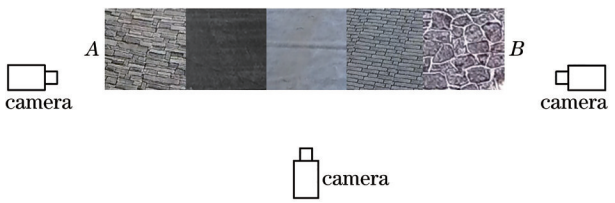

图3　实验环境与行走路线

Fig. 3　Experimental environment and walking route

提取监控录制的视频，使用HMSTranscoder视频转换器对其进行解码，并对处理后的视频进行分帧，根据实验需要，设置每三帧提取一帧图像，大小为1920 pixel×1080 pixel，通过人工筛选除去重复冗余的图像数据，每类鞋型大约包含200～300张图像。为保证实验数据的真实性与多样性，数据集还包含了鞋部部分遮挡、鞋部运动模糊以及光照等因素干扰的图像集。实验数据共计3062张图像，共制作标签数据5723个，如表1所示。

表1　训练集与测试集数据分布情况

Table 1　Data distribution of training set and testing set

Name	Traning set		Testing set	
	Object	Image	Object	Image
Shoe	5154	2755	569	307

制作检测标签使用的软件为labelimg，在对实验所得图像数据进行手工标注时，尽可能准确框选鞋部的全貌，并尽量避免边界框包含到不必要的背景图像。标签格式设置为PASCAL VOC数据集的格式，标签名称设置为“shoe”。标注情况如图4所示。

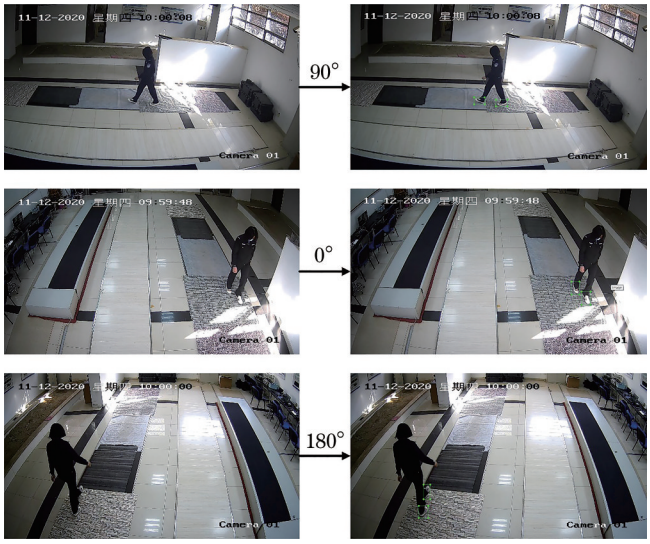

图 4　图像标注情况

Fig. 4　Image annotation

3.2　实验环境的配置

实验采用的硬件配置为 Intel(R) Xeon(R) E5-1650 v3 @ 3.50 GHz CPU，内存为 16 GB，NVIDIA GeForce GTX2080Ti GPU，显存为 11 GB；软件配置为 Windows 10 操作系统，CUDA 10.0 GPU 并行计算库，Pytorch 1.2.0 版本的深度学习框架。

整个实验训练过程进行 150 个 epoch，由于主干特征提取网络是通用的，为了加快训练速度，前 50 个 epoch 采用冻结网络的方式进行训练，学习率设置为 0.001，之后 100 个 epoch 的学习率设置为 0.0001，均使用 Adam 优化器。

3.3　模型参数设置

先验框的大小与形状会对模型训练的效果产生很大影响，准确的先验框能够更好地反映目标的特征，从而提高模型检测的精度。同时，先验框数量的设置也要权衡模型的召回率与检测效率，避免计算量过大增加模型的运行成本。本文通过 K-means 聚类的方法进行统计分析，从 1 开始设置聚类中心的个数，通过聚类运算分别获得了不同尺寸的先验框。实验过程发现，在聚类中心个数设置为 3 以内时，得到的先验框具有显著差异，有良好的代表性，但是当聚类中心个数增加到 4 以后，会发生冗余，得到尺寸相近的先验框。为了提高检测效率，精简模型参数，将模型的先验框数量设置为 3，模型输入图像的通道数为(608，608)，先验框尺度相应设置为(11，23)，(17，40)，(21，23)。

实验均采用 Mosaic 数据增强的方式，通过随机选用四张图片进行缩放、变化色域、拼接等方式，丰富了检测的数据集，同时也增强了模型的鲁棒性。此外，通过直接计算四张图片的数据，使得在训练过程中不需要设置很大的 batch size，也能够在 GPU 有限的情况下得到理想的检测效果。

3.4　评价指标

在目标检测的领域内，评价模型的检测效果通常需要计算精确度(precision)、召回率(recall)、平均精度 AP(average precision)，从而进一步计算 mAP(mean average precision)。AP 值为模型精确度与召回率绘制的 P-R 曲线与坐标轴所覆盖的面积，由于本文实验检测的种类只有鞋类一种，因此其 mAP 值与 AP 值相等。检测模型的大小通常通过其参数量(params)来表示。

3.5　实验结果分析

为了探究模型的检测层、多尺度特征融合以及空间金字塔池化结构对检测性能的影响，本文设置了不同的结构模型，通过实验确定检测效果最理想的改进方案。

3.5.1　模型检测层对检测性能的影响

为了探究不同检测层对模型检测性能的影响，实验根据 YOLOv4 模型主干特征提取网络 CSPdarknet53 的特点，一共设置了四组不同深度的主干网络。分别选用网络的 2、4、8、16 倍下采样层做检测层，以探究选用不同检测层对模型检测性能的影响，由浅到深分别对应 YOLOv4_A，YOLOv4_B，YOLOv4_C 与 YOLOv4_D 模型，如图 5 所示。

结果可知，YOLOv4_C 模型的 mAP 值达到 94.29%，表现最佳，如表 2 所示。原因是，针对小目

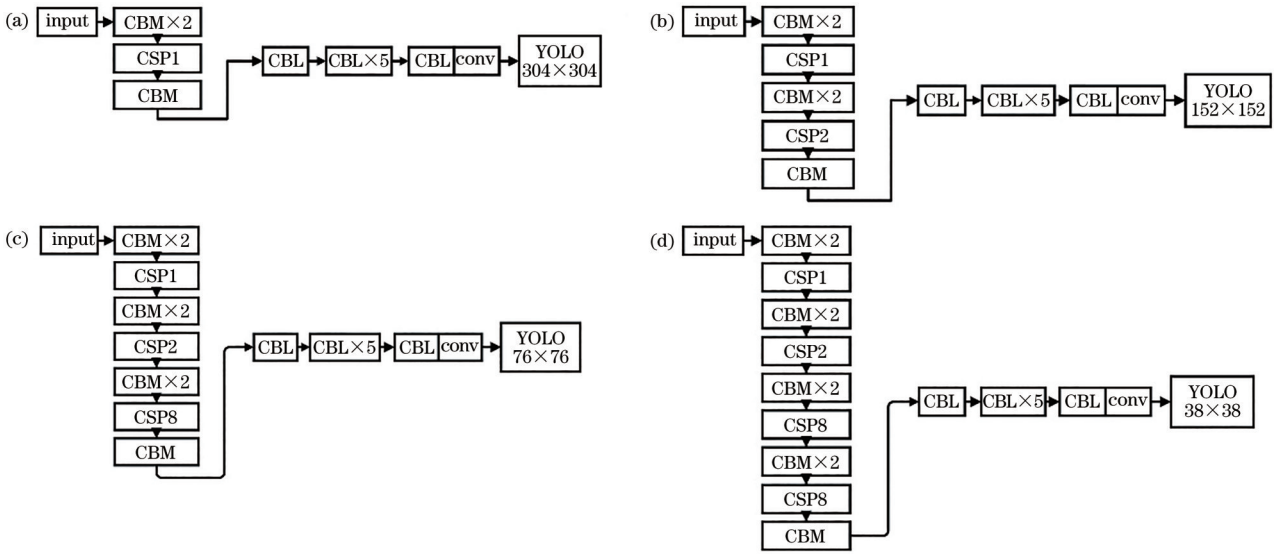

图 5　模型设置情况。(a) YOLOv4_A;(b) YOLOv4_B;(c) YOLOv4_C;(d) YOLOv4_D

Fig. 5　Model setting. (a) YOLOv4_A; (b) YOLOv4_B; (c) YOLOv4_C; (d) YOLOv4_D

表 2　实验结果

Table 2　Experimental results

Model type	mAP /%	Params /M
YOLOv4_A	69.04	0.51
YOLOv4_B	92.49	2.03
YOLOv4_C	94.29	11.81
YOLOv4_D	93.98	50.72

标的检测,当检测层较浅时,虽然网格更加细密,但是过浅的主干网络无法有效学习到目标的特征信息,从而导致检测效果变差;而在检测层数过深时,不利于获取到图像的细节信息,且会增加计算量,抑制模型的检测效果。综合考虑,选用 8 倍下采样层完成行人鞋部的检测工作。

3.5.2　特征融合对模型性能的影响

在确定模型的检测层后,为了尽量精简模型,实验仅保留了网络的 16 倍下采样层,在 neck 部分通过上采样与检测层进行特征融合,得到 YOLOv4_E 模型,如图 6 所示。

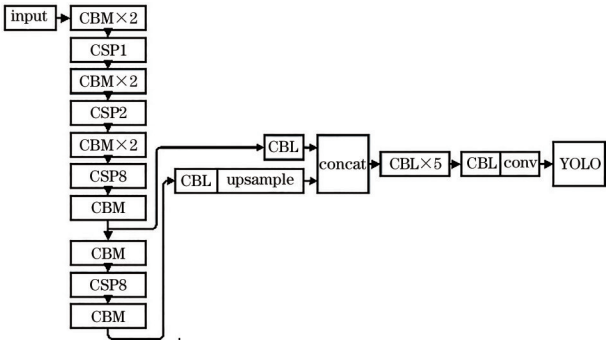

图 6　YOLOv4_E 模型

Fig. 6　YOLOv4_E model

YOLOv4_E 相对于 YOLOv4_C 增加了一条新的分支用于特征融合,其 mAP 值提高了 2.02%,如表 3 所示。原因是,深层网络感受野较大,能够提取到更丰富的语义特征,但是小目标在深层网络下的特征图太小,无法有效辨别,将浅层网络与深层网络进行多尺度特征融合,可以结合浅层网络的细节特征与深层网络的语义特征,使模型具有更好的描述性,从而增强模型对小目标的检测性能。

表 3　实验结果

Table 3　Experimental results

Model type	mAP /%	Params /M
YOLOv4_C	94.29	11.81
YOLOv4_E	96.31	38.81

3.5.3　空间金字塔池化结构对模型性能的影响

为了探究 SPP 结构对模型性能的影响,在 YOLOv4_E 的基础上,在 8、16 倍下采样层添加 SPP 模块形成模型 YOLOv4_F,在 16 倍下采样层添加 SPP 模块得到模型 YOLOv4_G,如图 7 所示。同时对 SPP 结构内的池化内核数与池化尺度进行调整,以得到适宜于小目标检测的池化方案。实验在 SPP 结构内使用 1×1,3×3,5×5 池化内核时命名为 SPP_1,使用 1×1,5×5,7×7 池化内核时命名为 SPP_2,使用 1×1,3×3,5×5,7×7 池化内核时命名为 SPP_3,使用 1×1,5×5,7×7,9×9 池化内核时命名为 SPP_4。

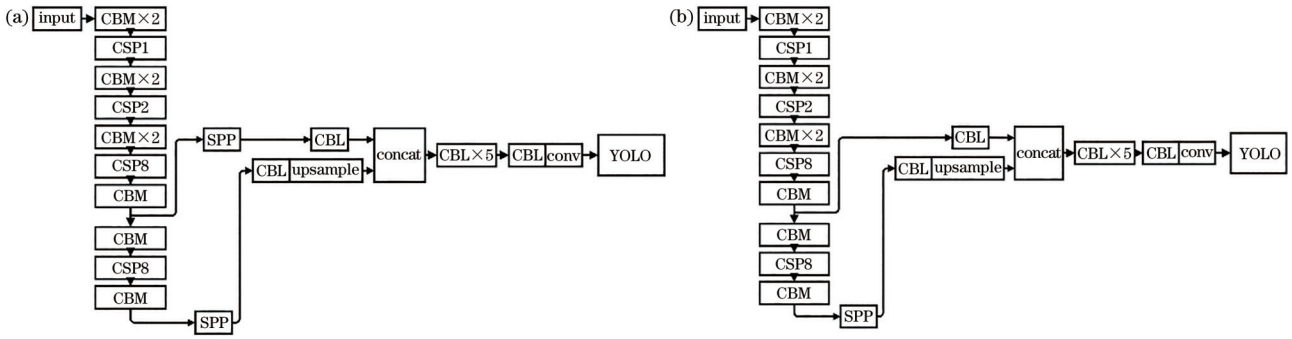

图 7　模型设置情况。(a) YOLOv4_F;(b) YOLOv4_G

Fig. 7　Model setting. (a) YOLOv4_F; (b) YOLOv4_G

结果发现,YOLOv4_F 与 YOLOv4_E 相比,虽然 neck 部分在 8,16 倍下采样层均增加了金字塔池化结构,但是其 mAP 值下降了 0.30%,而 YOLOv4_G 在只有 16 倍下采样层增加金字塔池化结构时,其 mAP 值要比 YOLOv4_E 的提高 0.40%,如表 4 所示。原因是,SPP 结构在 16 倍下采样层时,通过不同大小的池化结构,形成了多个层次信息,并通过相互融合增强学习能力,有效增加了图像特征的学习范围,可以较好地分离上下文特征,帮助模型学习到更多的信息,从而通过上采样与 8 倍下采样层特征融合,得到更好的检测结果;而将该结构添加到 8 倍下采样层时,反而增加了检测层的深度,并且最大池化的操作也会使鞋部细节特征流失,不利于模型浅层网络学习图像的位置、纹理等特征,最终导致模型对小目标检测性能下降。通过改进 SPP 结构内的池化核,可以有效增强检测效果。使用 SPP_1、SPP_2 结构均提升了检测表现,说明使用较小的 3×3,5×5,7×7 池化核虽然感受野减小,但是在减少鞋部特征流失的同时,可避免过大的感受野获取到的无关信息造成干扰,从而有效提升了检测性能。最终使用

表 4　实验结果

Table 4　Experimental results

Model type	mAP /%	Params /M
YOLOv4_E	96.31	38.81
YOLOv4_F	96.01	39.93
YOLOv4_G(SPP)	96.71	39.56
YOLOv4_G(SPP_1)	96.98	39.31
YOLOv4_G(SPP_2)	97.33	39.31
YOLOv4_G(SPP_3)	97.93	39.56
YOLOv4_G(SPP_4)	96.94	39.56

SPP_3 代替原 SPP 结构,mAP 提升了 1.22%。而 SPP_4 与 SPP_2 相比,增加了 9×9 的池化核之后,检测效果变弱,这是因为检测目标本身较小,且在 16 倍下采样层时,图像特征经过进一步的压缩,已经丢失了一定量的信息,如果使用大尺度的池化核,则会造成更多鞋部特征信息的损失,从而影响检测效果。

综合可得,YOLOv4_G(SPP_3)模型实验表现最佳,将该模型命名为 YOLOv4_shoe,该模型训练过程的损失函数变化曲线以及召回率(P-R)曲线如图 8 所示。

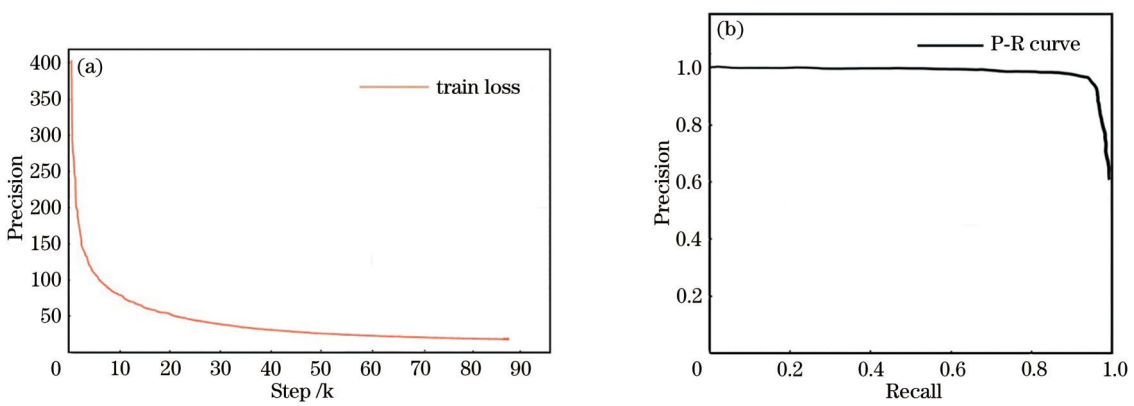

图 8　YOLOv4_shoe 的训练情况。(a)损失函数变化曲线;(b) P-R 曲线

Fig. 8　Training of YOLOv4_shoe. (a) Loss function curve; (b) P-R curve

3.5.4　与其他检测模型的精度对比

表5为不同模型检测不同监控视角下行人鞋部目标的结果。YOLOv4_shoe算法的检测速度较SSD，YOLOv3模型较低，可能是由于通道数不同及检测思路不同等原因所致，但是YOLOv4_shoe模型的检测精度及训练参数量均优于其他算法。因此，YOLOv4_shoe模型针对鞋部小目标的检测效果更为理想。

表5　不同算法的检测性能对比

Table 5　Detection performance comparison of different algorithms

Network structure	mAP /%	Params /M	FPS /(f·s^{-1})
SSD	86.43	90.60	24.33
YOLOv3	91.92	234.98	29.03
YOLOv4	95.86	244.30	14.80
YOLOv4_shoe	97.93	39.56	23.71

3.5.5　模型检测结果

YOLOv4与YOLOv4_shoe模型在测试集上检测效果如图9所示。相比于YOLOv4模型，改进后的模型置信度更高，在轻微遮挡、运动模糊以及行人多方向行进等干扰因素下，模型的检测效果均较为理想。

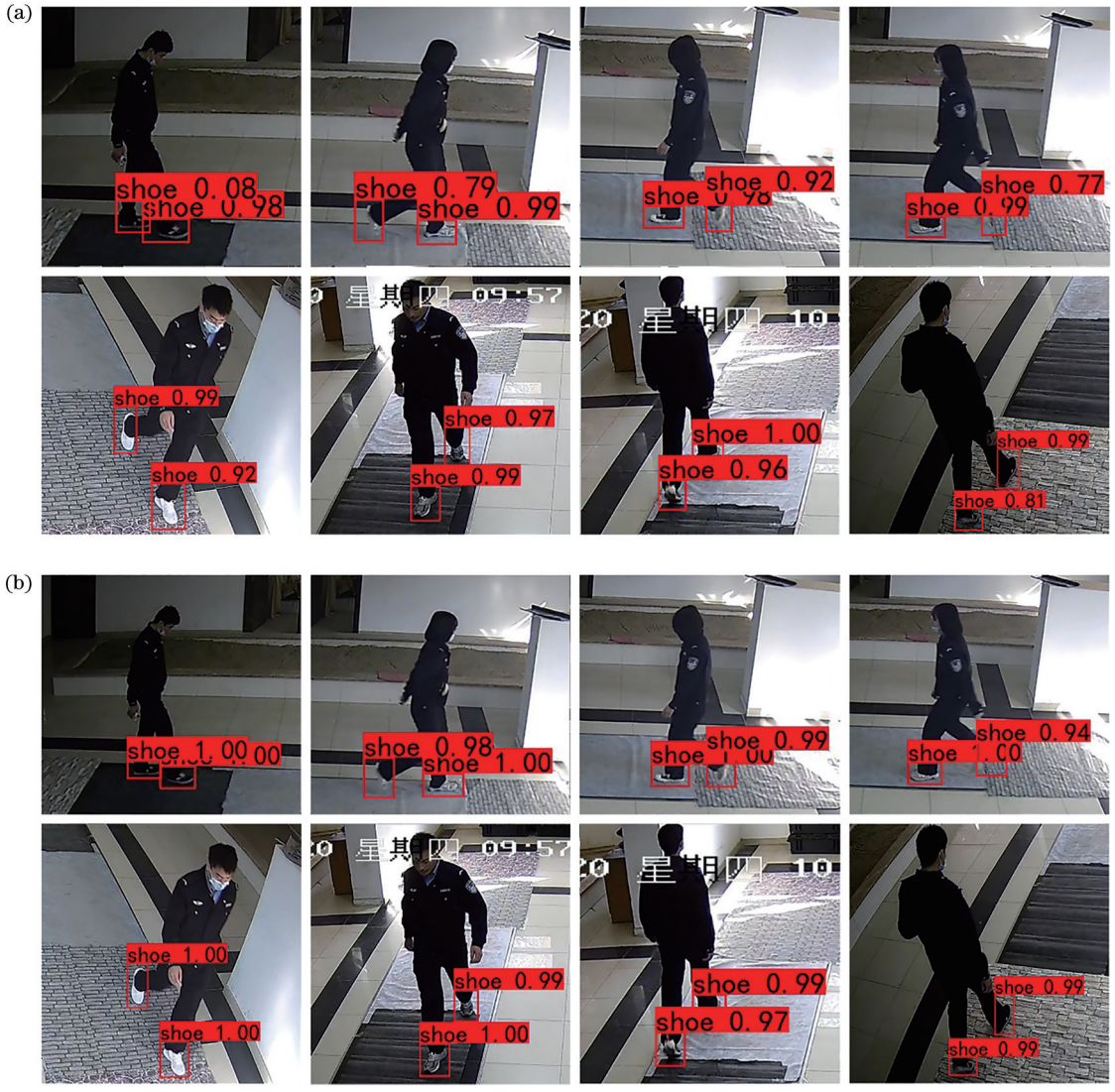

图9　检测结果。(a) YOLOv4；(b) YOLOv4_shoe

Fig. 9　Detection results. (a) YOLOv4; (b) YOLOv4_shoe

YOLOv4_shoe模型大部分情况下均能准确检测目标，其主要误差来源如图10所示。图10(a)为由于运动导致的图像模糊情况；图10(b)为运动过程中由于裤子、鞋子等遮挡的情况；图10(c)为由于光线不均导致的检测误差的情况；图10(d)为行人鞋部与行走背景混同造成影响的情况。

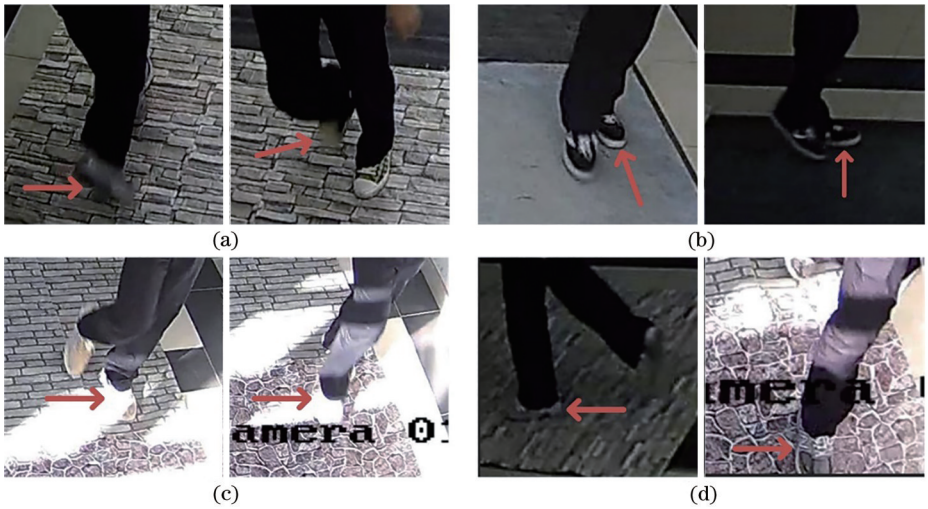

图10　检测误差情况。(a)运动模糊；(b)遮挡；(c)光照不均；(d)图案混淆

Fig. 10　Detection error condition. (a) Motion blur; (b) cover; (c) uneven illumination; (d) confusion of patterns

4　结　　论

基于YOLOv4算法改进后得到的YOLOv4_shoe模型，通过重新设置先验框尺寸、减少先验框数量；降低主干网络的层数，删去16、32倍下采样层对应的检测头；使用多尺度特征融合、调整空间金字塔池化结构这些途径，可以有效增强网络对监控下行人鞋部目标的检测性能，得到了一个轻量高效的目标检测模型。

YOLOv4_shoe模型在IoU阈值为0.5的情况下检测效果较好，事实上，实战中鞋类检索工作对鞋部检测的定位准确性要求较高，但是现阶段实验网络结构较为简单，实验数据缺乏多人、行进速度加快等干扰因素下的数据，具有一定的局限性。下一步工作需要继续改进优化网络，提高IoU值阈值较高时的检测精度，构建复杂场景下高效稳定的鞋部检测网络模型。

参　考　文　献

[1] Yuan C P, Yu S W. Preliminary study on the application of footprint analysis in video investigation [J]. Guangdong Gongan Keji, 2017, 25(2): 61-63, 74.
袁楚平, 余尚伟. 足迹分析在视频侦查工作中的运用初探[J]. 广东公安科技, 2017, 25(2): 61-63, 74.

[2] Nong D S. Based on the scene footprint, expand video investigation[J]. Legal System and Society, 2015(22): 255-256.
农达史. 立足现场足迹、拓展视频侦查[J]. 法制与社会, 2015(22): 255-256.

[3] Yang M J, Tang Y Q, Jiang X J. Novel shoe type recognition method based on convolutional neural network[J]. Laser & Optoelectronics Progress, 2019, 56(19): 191505.
杨孟京, 唐云祁, 姜晓佳. 基于卷积神经网络的鞋型识别方法[J]. 激光与光电子学进展, 2019, 56(19): 191505.

[4] Geng P Z, Yang Z X, Zhang J J, et al. Pedestrian shoes detection algorithm based on SSD[J]. Laser & Optoelectronics Progress, 2021, 58(6): 061009.
耿鹏志, 杨智雄, 张家钧, 等. 基于SSD的行人鞋子检测算法[J]. 激光与光电子学进展, 2021, 58(6): 061009.

[5] Nan X H, Ding L. Review of typical target detection algorithms for deep learning[J]. Application Research of Computers, 2020, 37(S2): 15-21.
南晓虎, 丁雷. 深度学习的典型目标检测算法综述[J]. 计算机应用研究, 2020, 37(S2): 15-21.

[6] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR 2001, December 8-14, 2001, Kauai, HI, USA. New York: IEEE Press, 2001: 7176899.

[7] Viola P, Jones M J. Robust real-time face detection

[J]. International Journal of Computer Vision, 2004, 57(2): 137-154.

[8] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), June 20-25, 2005, San Diego, CA, USA. New York: IEEE Press, 2005: 886-893.

[9] Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model[C]//2008 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2008, Anchorage, AK, USA. New York: IEEE Press, 2008: 10139902.

[10] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE Press, 2014: 580-587.

[11] Girshick R. Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 1440-1448.

[12] 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.

[13] He K M, Gkioxari G, Dollár P, et al. Mask R-CNN [C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2980-2988.

[14] Redmon J, Divvala S, Girshick R, et al. You only look once: unified, real-time object detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 779-788.

[15] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 6517-6525.

[16] Redmon J, Farhadi A. Yolov3: an incremental improvement[EB/OL]. (2018-04-01) [2021-02-01]. https://arxiv.org/abs/1804.00276.

[17] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: optimal speed and accuracy of object detection[EB/OL]. (2020-04-23) [2021-02-01]. https://arxiv.org/abs/2004.10934v1.

[18] Liu W, Anguelov D, Erhan D, et al. SSD: single shot MultiBox detector[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9905: 21-37.

[19] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision (ICCV), October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[20] He K M, Zhang X Y, Ren S Q, 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.

[21] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 936-944.

[22] Liu S, Qi L, Qin H F, et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 8759-8768.

[23] Zheng Z H, Wang P, Liu W, et al. Distance-IoU loss: faster and better learning for bounding box regression[EB/OL]. (2019-11-19)[2021-02-01]. https://arxiv.org/abs/1911.08287v1.

[24] Li B, Wang C, Wu J, et al. Surface defect detection of aeroengine components based on improved YOLOv4 algorithm[J]. Laser & Optoelectronics Progress, 2021, 58(14): 1415004.
李彬, 汪诚, 吴静, 等. 改进YOLOv4算法的航空发动机部件表面缺陷检测[J]. 激光与光电子学进展, 2021, 58(14): 1415004.

[25] Liu Y, Zhan Y W. Survey of small object detection algorithms based on deep learning[J]. Computer Engineering and Applications, 2021, 57(2): 37-48.
刘洋, 战荫伟. 基于深度学习的小目标检测算法综述[J]. 计算机工程与应用, 2021, 57(2): 37-48.

[26] Gao Y, Xiao D. Small object detection algorithm based on multi-feature fusion[J]. Computer Engineering and Design, 2020, 41(7): 1905-1909.
高杨, 肖迪. 基于多层特征融合的小目标检测算法[J]. 计算机工程与设计, 2020, 41(7): 1905-1909.