

DOI:10.16136/j.joel.2023.07.0442

基于轻量化YOLOv3的带钢表面缺陷检测方法

马千文¹, 刘国华^{1,2*}

(1. 天津工业大学 机械工程学院, 天津 300387; 2. 天津市现代机电装备技术重点实验室, 天津 300387)

摘要: 缺陷检测是带钢生产过程中不可缺少的工序, 现有检测方法普遍存在检测精度较低、实时性差等问题。为解决上述问题, 本文提出了一种基于轻量化YOLOv3的快速缺陷检测方法。MobileNetv2作为主干网络并用两个尺度的特征图进行输出, 保证了网络模型的轻量化; 将改进后的注意力模块融合进特征金字塔网络(feature pyramid network, FPN), 同时结合空间金字塔池化模块(spatial pyramid pooling, SPP), 以提高算法对缺陷的学习能力; 使用K均值聚类算法获得更优的先验框, 并且使用CIoU(complete-intersection over union)对损失函数进行优化, 进一步提升网络性能。提出的方法在带钢缺陷数据集上检测速度为70.8 FPS; 模型参数量为7.1 MB, 仅为YOLOv3的3.02%。实验结果表明本文所提方法能够在保证精度的同时实现对缺陷的快速检测, 具有良好的生产线部署能力。

关键词: 金属带钢; 缺陷检测; 深度学习; 卷积神经网络; YOLOv3

中图分类号: TP391.4 **文献标识码:** A **文章编号:** 1005-0086(2023)07-0743-09

Surface defect detection method of strip steel based on lightweight YOLOv3

MA Qianwen¹, LIU Guohua^{1,2*}

(1. School of Mechanical Engineering, Tiangong University, Tianjin 300387, China; 2. Advanced Mechatronics Equipment Technology Tianjin Major Laboratory, Tianjin 300387, China)

Abstract: Defect detection is an indispensable process in the strip steel production process, and existing inspection methods generally have problems such as low detection accuracy and poor real-time performance. To solve the above problems, a fast defect detection method based on lightweight YOLOv3 is proposed in this paper. MobileNetv2 is used as the backbone network and output with two scales of feature maps, so that the lightweight of the network model is guaranteed; the improved attention module is fused into the feature pyramid network (FPN) and the network is combined with the spatial pyramid pooling (SPP) to improve the learning ability of the algorithm for defects; the K-means mean clustering algorithm is used to obtain a better anchor box, and the complete-intersection over union (CIoU) is used to optimize the loss function to further improve the network performance. The proposed method has a detection speed of 70.8 FPS on the strip steel defect dataset; the number of model parameters is 7.1 MB, which is only 3.02% of YOLOv3. Experiments show that the proposed method can achieve rapid detection of defects while ensuring accuracy, and has good production line deployment capabilities.

Key words: metal strip steel; defect detection; deep learning; convolutional neural network; YOLOv3

0 引言

带钢是冶金工业的重要产品之一, 在日常生活以及工业生产中应用十分广泛^[1], 其表面质量是衡量产品整体质量好坏的关键因素。由于制作

工艺以及外部原因, 其成品会出现孔洞、裂纹、折痕、轧坑以及氧化铁皮等缺陷。上述缺陷不仅会影响产品外观, 还会成为应力集中点、受力薄弱点、性能突变点、疲劳损伤等问题的策源地, 从而降低产品的耐磨性、耐腐蚀性、机械强度以及其他

* E-mail: liuguohua@tiangong.edu.cn

收稿日期: 2022-06-16 修订日期: 2022-09-12

基金项目: 天津市科技计划项目(21YFJCYS00080)资助项目

性能,因此带钢表面缺陷检测就成了其生产制造过程中不可缺少的步骤。近几年来,表面缺陷检测技术得到越来越多人的关注^[2],为了确保最终产品的质量,有必要在带钢的生产过程中进行高效、实时的缺陷检测。

随着科学技术的发展,人们对检测标准要求越来越高,机器视觉检测技术正在逐步地应用于各领域之中。方春华等^[3]使用 Canny 算法计算缺陷边缘的梯度和幅值,增加了缺陷边缘的清晰度,能够对所检测物体实现较高的检测精度;AGH-DAM 等^[4]使用二值化处理来对缺陷的特征进行提取,然后构建分类器对金属带钢表面的缺陷进行识别,相对于传统非视觉方法来说具有更好的实时性;李俊峰等^[5]利用高斯双方向导数滤波器对缺陷图像进行处理,对亮点、压伤等缺陷实现了较好的检测效果。虽然上述传统图像处理方法能够满足表面缺陷检测基本的实时性和精确度要求,但是对光照强度、背景复杂度等外部条件要求较高。面对不同背景需要调整算法的阈值参数,面对不同的缺陷特征需要修改甚至编写不同的检测算法,难以使用一套算法模型对不同类型缺陷进行检测,算法泛化能力和适应能力较弱。

深度学习凭借其在图像处理方面的优异表现,为实现带钢表面缺陷检测提供了一个新思路^[6]。深度学习中的卷积神经网络可以对缺陷的特征进行有效学习和提取,既可以在低层神经网络中获得精确的位置信息,又可以在高层神经网络中获得丰富的语义信息。且深度学习方法的强泛化能力可以满足工业上对鲁棒性及高效性的要求,使其可以在不同外部环境下使用一套算法模型来对不同类型的表面缺陷进行检测。谷静等^[7]提出了一种改进的深度学习缺陷检测方法,使用

膨胀卷积来增加感受野,从而提升了检测的精度;方叶祥等^[8]对 YOLOv3 进行改进,在图像输入端进行直方图均衡化处理,然后对缺陷进行定位,达到了较好的检测效果。但是以上方法只注重提高检测的精度,忽略了检测的速度和实时性,不适合部署到实际工业生产线上。

针对上述方法存在的问题,本文对 YOLOv3 进行改进,提出了一种基于轻量化 YOLOv3 的快速缺陷检测方法,具体工作如下:

1) 针对目前带钢表面缺陷检测领域存在鲁棒性较差、检测速度较慢、无法使用一套算法完成对不同缺陷的检测任务等问题,本文提出了一种基于深度学习的快速缺陷检测方法。为保证网络模型的轻量化,将原网络中 Darknet53 更换为速度更快的 MobileNetv2^[9]并减少了一个尺度的特征图输出,提高了实时检测能力以及移动端部署能力。

2) 本文分析了当前特征金字塔网络(feature pyramid network, FPN)^[10]存在的一些问题,不局限于从特征融合的角度进行改进,而是充分挖掘其潜在特征,根据不同层的特征图对信息的敏感程度不同,对 CBAM(convolutional block attention module)^[11]进行改进并将其融合进入 FPN。

1 轻量化 YOLOv3 模型

原始 YOLOv3 模型^[12]虽然能够保证精度和速度上的相对均衡,但是由于深度学习网络模型参数量较大,其结构从硬件部署角度来说仍然相对复杂。本文所提方法主要在保证精度的前提下减少其参数量,提升检测速度,图 1 为本文所提方法的总网络框架,其中卷积模块的解释如图 2 所示。

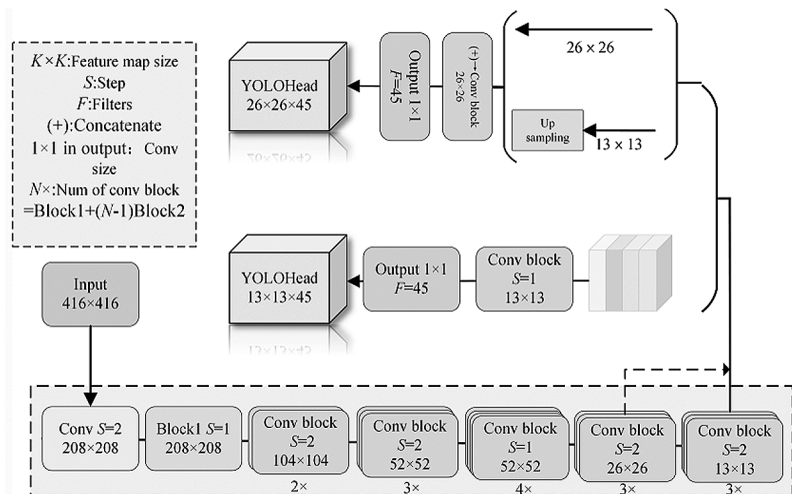


图 1 轻量化 YOLOv3 模型

Fig. 1 Lightweight YOLOv3 model

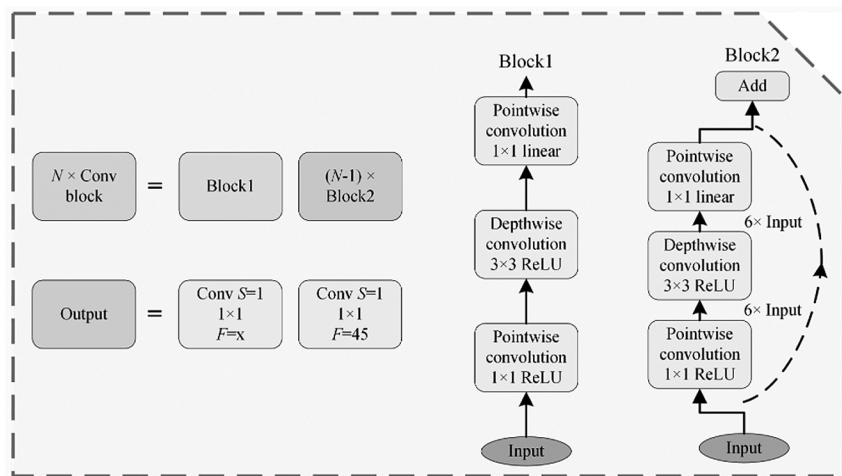


图 2 网络结构中的卷积模块

Fig. 2 Convolutional modules in the network structure

1.1 模型的轻量化操作

为了减少模型的计算量,将原 YOLOv3 模型中的主干特征提取网络更换为 MobileNetv2,并且减少了一个尺度特征图的输出。MobileNetv2 中大量使用了图 2 中的 Block 卷积块,其中残差结构 Block2 不同于 ResNet^[13] 中的残差结构,ResNet 为了减少计算量,在其残差结构中进行了大量通道数的压缩,此种做法在一定程度上限制了网络的特征提取能力。而 Block2 是一种倒置残差结构,首先通过第一个逐点卷积(pointwise convolution)将输入端通道数进行了 6 倍扩张,然后通过深度卷积(depthwise convolution)进行特征提取,最后通过输出端的逐点卷积压缩通道数,使得输入端和输出端通道数保持不变。因为在压缩通道时,ReLU 激活函数会破坏特征,影响网络效果,而 Linear 为线性函数,因此在输出端,ReLU 函数被替换为 Linear 函数,来避免特征损失。MobileNet 的核心是标准卷积被替换为深度可分离卷积块^[14],在保证检测性能的前提下降低网络计算成本。

如图 3 所示,对于标准卷积,假设输入图像尺寸为 $D_f \times D_f \times A$,每个卷积核计算量为 $D_k \times D_k \times A$,输出特征图尺寸为 $D_f \times D_f \times B$,因为标准卷积的每一个卷积核都需要对所有输入通道的特征图进行计算,所以其计算参数量为 $D_k \times D_k \times A \times B$ 。

深度可分离卷积见图 4,由两部分组成,如图 4(a)所示的深度卷积和图 4(b)所示的逐点卷积。深度卷积和标准卷积的最大区别为深度卷积的每个卷积核仅需对一个输入通道特征图进行计算,每个卷

积核计算量为 $D_k \times D_k \times 1$,当输入图像尺寸为 $D_f \times D_f \times A$,输出特征图尺寸为 $D_f \times D_f \times A$ 时,在深度卷积部分其计算量为 $D_k \times D_k \times 1 \times A$;逐点卷积的主要任务为获得期望的输出通道数,其输入端为深度卷积的输出,每个卷积核计算量为 $1 \times 1 \times A$,当输出为 $D_f \times D_f \times B$ 时,其计算量为 $1 \times 1 \times A \times B$ 。综上,深度可分离卷积的总计算量为 $D_k \times D_k \times 1 \times A + 1 \times 1 \times A \times B = D_k \times D_k \times 1 + A \times B$ 。

深度可分离卷积和标准卷积完成一个卷积过程的参数量对比如式(1)所示:

$$\frac{D_k \times D_k \times A + A \times B}{D_k \times D_k \times A \times B} = \frac{1}{B} + \frac{1}{D_k \times D_k}, \quad (1)$$

式中, D_k 为卷积核尺寸, A 为输入特征图通道数, B 为输出特征图通道数,由公式可看出深度可分离卷积减少了计算的参数量。

如表 1 所示为原 YOLOv3 网络与改进之后网络的模型对比,其中 YOLOv3 的主干网络为 Darknet53, Improved 1 和 Improved 2 的主干网络均为 MobileNetv2。检测头数目(number of detection heads, NDH)代表输出的检测层个数;模型复杂度(BFLOPS)代表模型计算需要的十亿次浮点运算数目;模型参数量(model size)为模型训练生成的权重大小;FPS 为检测速度的单位。在经过网络模型的轻量化处理之后,模型参数量由 235 MB 变为 7.1 MB,仅为原 YOLOv3 的 1/33.1(3.02%);模型复杂度由 65.37 变为 3.049,为原 YOLOv3 的 1/21.44(4.66%),改进后的网络模型有效实现了模型的轻量化,具备了更强的移动端部署能力,更易部署于带钢表面缺陷检测生产线中。

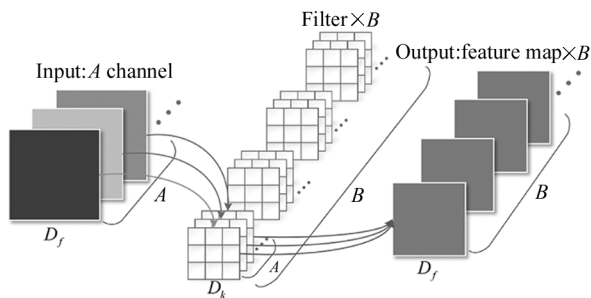


图3 标准卷积示意图

Fig. 3 Diagram of standard convolution

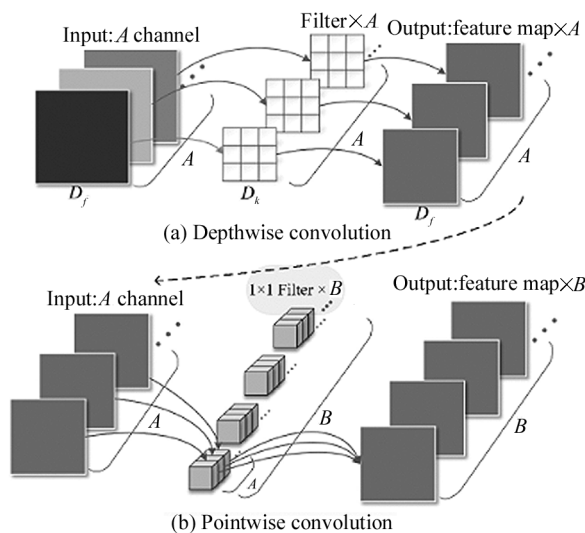


图4 深度可分离卷积示意图

Fig. 4 Diagram of depthwise separable convolution

表1 轻量化模型对比

Tab. 1 Comparison of lightweight models

Model	NDH	BFLOPS	Model size/MB	Detection speed/FPS
YOLOv3	3	65.370	235	26.5
Improved 1	3	3.918	7.73	66.1
Improved 2	2	3.049	7.10	70.8

1.2 改进的特征金字塔

FPN 结构通过融合不同尺度的特征图有效利用了主干网络提取的缺陷特征,目前已经成为目标检测最常用的结构之一。很多学者对其做出了各种各样的改进,然而仍然存在一些不足,主要是没有充分挖掘原 FPN 的潜在能力。不同大小的特征图对通道和空间的敏感程度是不同的,低层的特征图语义信息比较少、但是含有丰富的位置信息^[15],更加关注于空间特征;高层的特征图位置信息比较粗糙、但是含有丰富的语义信息,更加关注于通道特征。CBAM 是一个同时考虑到空间注意力特征和通道注意力特征的注意力模块。分别在空间上和通道上进

行了注意力的加权,不仅控制了网络模型的计算量和参数量,而且能够作为即插即用的模块插入到现有网络中。但是,其传递方式为串联传递,忽略了不同尺度特征图对空间特征和通道特征的关注程度不同这一特点。针对带钢表面缺陷不易提取、特征不明显等工业现状以及上述问题,并结合 CBAM 的轻量化特点,将 CBAM 进行改进并融合到 FPN 中。结合低层特征图和高层特征图各自的特点,赋予其不同的注意力模块优先级。对于低层特征图赋予其更多的空间注意力特征,对于高层特征图赋予其更多的通道注意力特征。具体结构如图 5 所示。

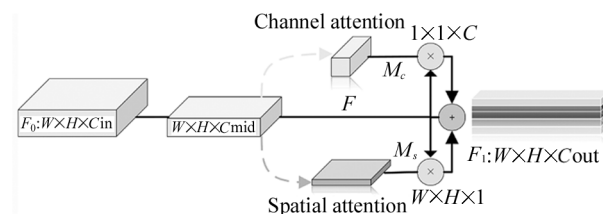


图5 融合不同注意力加权的 FPN

Fig. 5 Fusion of FPN with different attention weights

图 5 中,输入特征图为 F_0 ,通道注意力模块和空间注意力模块计算式如下所示:

$$M_c(F) = \sigma(MLP(AvgPool(F) + MLP(MaxPool(F)))) \quad (2)$$

$$M_s(F) = \sigma(f^{7 \times 7}(Concat(AvgPool(F), MaxPool(F)))) \quad (3)$$

式中, F 为中间过程中的特征图, $M_c(F)$ 为通道注意力模块的输出, $M_s(F)$ 为空间注意力模块的输出, σ 为 sigmoid 函数, MLP 表示由两个全连接层和 ReLU 激活函数组成的多层感知机, $f^{7 \times 7}$ 表示卷积核为 7×7 的卷积运算。FPN 的整体结构如图 6 所示,其中输入端 F_0 为主干特征提取网络,由于模型的轻量化缩减了一个尺度特征图的输出,所以仅对 13×13 和 26×26 两个尺度的特征图进行融合与输出。首

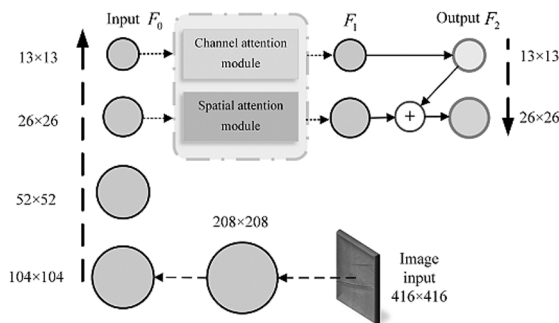


图6 FPN 示意图

Fig. 6 Overall diagram of FPN

先 13×13 和 26×26 的特征图经过注意力模块处理, 13×13 的特征图传入检测头中进行预测, 然后网络向下传递, 经过上采样之后与 26×26 的特征图进行融合并输出。

1.3 应用空间金字塔池化模块

带钢工业生产线受作业环境影响, 缺陷特征相对不易提取。而空间金字塔池化模块 (spatial pyramid pooling, SPP)^[16] 中的多元感受野可以实现多尺度局部特征信息和全局特征信息的有效融合, 提供更加丰富的特征信息, 有效利用带钢缺陷特征; 且可以在几乎不增加计算量的情况下, 实现特征的融合, 不会对检测速度造成影响。利用 SPP 的上述特点, 结合带钢生产线实时检测需要, 本文在检测算法中应用 SPP 模块。如图 7 所示, 输入端为 $13 \times 13 \times 160$ 的特征图, 使用 3 个不同尺度的最大池化层对其进行处理, 池化核的大小分别为 5×5 , 9×9 , 13×13 , 步长设置为 1。然后将池化之后的特征图与原始特征图进行 concatenate 操作。模型中将最大池化核尺寸设置为 13×13 , 与输入特征图尺寸相同, 池化之后均输出尺寸为 $13 \times 13 \times 160$ 的特征图, 从而实现全局特征与三个池化核输出的局部特征相互融合, 有效增大了感受野。经过 SPP 模块的一系列池化、concatenate 操作, 输入端 $13 \times 13 \times 160$ 的特征图变为 $13 \times 13 \times 640$ 的特征图。

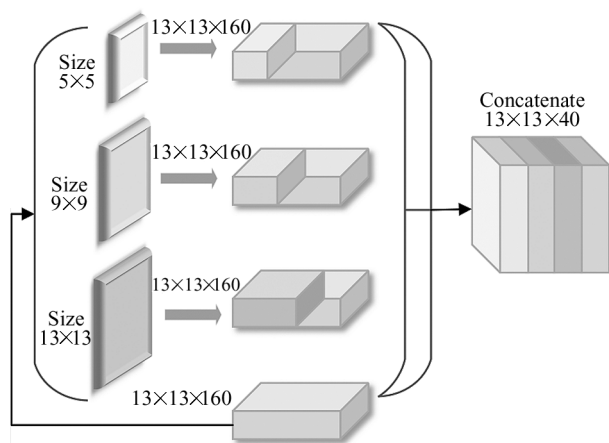


图7 空间金字塔池化模块

Fig. 7 Spatial pyramid pooling

1.4 损失函数优化

交并比 (intersection over union, IoU) 是目标检测算法中非常重要的指标, 意为“预测的边框 (C)”与“真实的边框 (D)”交集和并集的比值, 其计算式如下所示:

$$IoU = \frac{|C \cap D|}{|C \cup D|}, \quad (4)$$

虽然 IoU 能够较为简单地计算 loss, 但是存在两个问题:

1) 如图 8(a) 所示, 当预测框和真实框不相交时, $IoU = 0$, 无法读取两个框的距离信息, 此时损失函数无法计算。

2) 如图 8(b) 和 8(c) 所示, 两种情况的 IoU 相同, 但此时无法利用 IoU 将其区分。

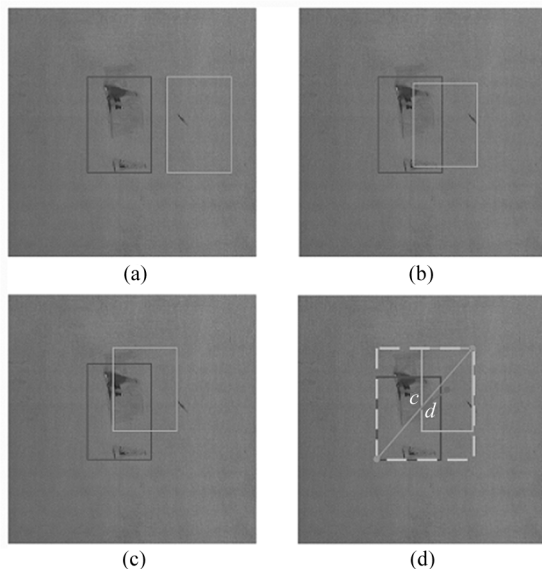


图8 真实框与预测框位置关系:

(a) 无交集; (b)(c) 交集相同; (d) 考虑多个几何因素

Fig. 8 Relationship between the ground truth bounding box and the predicted bounding box: (a) No intersection; (b)(c) Same intersection; (d) Considering multiple factors

为解决上述问题, 使用 $CIoU$ (complete-intersection over union)^[17] 作为损失函数, 如图 8(d) 所示, 增加外接矩形, 计算与外接矩形的差值, 解决真实框和预测框无接触或者交集相同的问题; 将边界框中心点距离考虑进来, 解决两框外接矩形相同的问题; 将边界框宽高比尺寸信息考虑进来, 解决宽高比不统一的问题。 $CIoU$ 将损失函数的 3 个重要几何因素: 重叠面积、中心点距离、长宽比都进行了考虑。 $CIoU$ 损失函数计算过程如式 (5) 所示:

$$CIoU_Loss = 1 - CIoU = 1 - \left(IoU - \frac{Distance_d^2}{Distance_c^2} - \frac{v^2}{(1 - IoU) + v} \right), \quad (5)$$

式中, $Distance_c$ 为对角线距离, $Distance_d$ 为两个中心点的欧氏距离。 v 为衡量长宽比一致性的参数, 其计算过程如式 (6) 所示:

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w^p}{h^p} \right), \quad (6)$$

式中, w^{gt} 和 h^{gt} 分别表示真实框的宽高, w^p 和 h^p 则分别表示预测框的宽高, $\arctan \frac{w}{h}$ 表示框对角线和高度之间的角度。

2 数据集部分

本文所用数据集为 GC10-DET, 为生产现场实地采集的带钢表面缺陷数据集, 共含有月牙形缺口(Cg)、异物(In)、轧坑(Rp)、焊缝(Wl)、折痕(Cr)、丝斑(Ss)、油斑(Os)、水斑(Ws)、冲孔(Pu)和腰折(Wf)这 10 种缺陷类型, 数据集中样本数量为 4 003 张, 缺陷图像如图 9 所示。

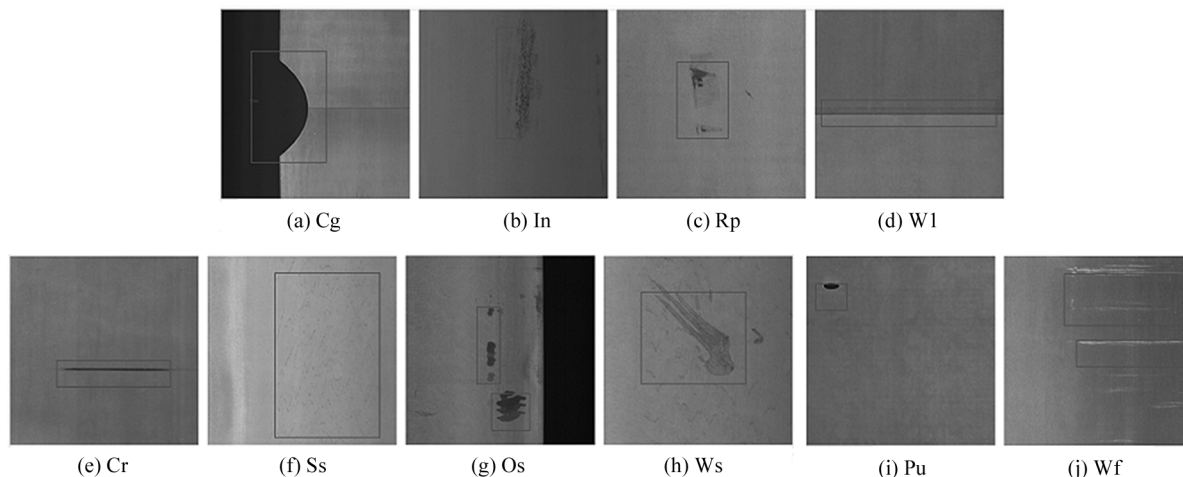


图 9 带钢表面缺陷图像

Fig. 9 Defect image on strip steel surface

的聚类结果如图 10 所示。

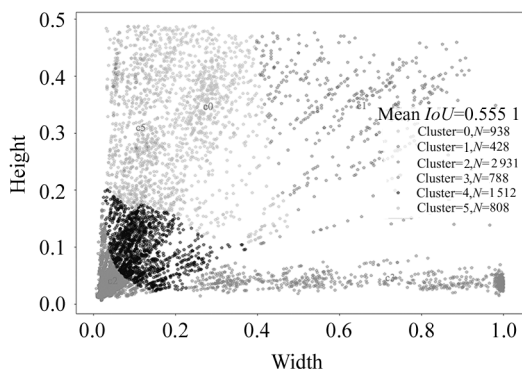


图 10 聚类结果

Fig. 10 The result of clustering

3 实验部分

此部分使用多个模型在带钢表面缺陷数据集中进行了多次实验, 得到了一系列的数据证明所提方

法在不同类型的缺陷具有不同的尺寸, 所以有必要在训练之前设置好适合带钢缺陷尺寸特点的先验框。不同于原网络使用默认的先验框, 本文使用 K 均值聚类算法根据 GC10-DET 数据集中带钢缺陷特点, 自动选取了 6 类出现频率高的边缘框尺寸作为其先验框。聚类过程如下: 1) 首先设定 6 个初始聚类中心; 2) 计算每个标注框和聚类中心点的距离, 将标注框分配给距离最近的聚类中心; 3) 所有标注框分配完后, 对每个簇重新计算聚类中心; 4) 循环执行 2)、3) 步直到聚类中心位置不再变化。聚类后的先验框尺寸分别为 (14, 29)、(47, 84)、(298, 34)、(43, 257)、(111, 290) 和 (268, 292), 读取数据集中样本信息以及标注信息, 使用聚类算法进行处理, 生成

法在金属带钢缺陷检测中的有效性。首先对实验环境进行介绍, 然后制定一系列评价指标进行多比例模型训练, 并且将本文所提方法和目前主流算法模型进行对比, 最后通过消融实验来验证所用方法在性能提升上的有效性。

3.1 环境平台搭建

本文实验基于 Windows10 系统进行, 处理器为 AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz, 内存为 16 GB, GPU 为 NVIDIA GEFORCE RTX3060。利用 CUDA11.0 和 cuDNN8.0.1 来加速 GPU 运算, 编译器使用 PyCharm 并且安装多个第三方库支持模型构建。

3.2 评价指标

在工业检测中, 检测的精度和速度尤为重要。实际检测过程中, 如果出现检测速度过慢、缺陷检测精度过低等问题, 会导致机械设备做出错误判断从

而影响工业生产线正常高效运行。

综上考虑,使用均值平均精确度(mean average precision, mAP)作为衡量缺陷检测算法检测性能的重要指标之一, mAP 也是目标检测算法中经常使用的评价指标,意为 n 类缺陷平均精度的平均值,计算式如下:

$$mAP = \frac{\sum_{i=1}^n AP_i}{n} \quad (7)$$

$Precision$ 为查准率,代表分类器预测正例的准确程度,计算式如下:

$$Precision = TP / (TP + FP) \quad (8)$$

$Recall$ 为查全率,代表分类器对正例的覆盖能力,计算式如下:

$$Recall = TP / (TP + FN) \quad (9)$$

式中, TP (true positives) 为正样本被正确识别为正样本的数目, FP (false positives) 为负样本被错误识别为正样本的数目, FN (false negatives) 为正样本被错误识别为负样本的数目,此外还有 TN (true negatives) 为负样本被正确识别为负样本的数目。检测速度关系到产品的生产效率,在工业生产线中尤为重要,因此,检测速度同样作为评价指标,单位 FPS 为每秒能处理的图像帧数。平均精度(average precision, AP),由查准率和查全率计算得出。将 $Precision$ 作为纵坐标,将 $Recall$ 作为横坐标,构成 PR 曲线($Precision-Recall$ 曲线),代表查准率和查全率的关系。查准率和查全率具有此消彼长的关系, AP 是 PR 曲线和 x 轴、 y 轴围成的面积,计算式如下:

$$AP = \int_0^1 p_{smooth}(r) dr \quad (10)$$

3.3 模型训练

在实验中发现,同一个模型设置不同的训练参数,训练结束之后的网络模型性能存在一定差异。为尽可能通过调参实现模型的最佳性能,同时又要考虑计算机硬件对参数的最大调整阈值限制,经过多次实验,将模型参数设置如下:输入图像原始尺寸设置为 416×416 , batch size 为 4,使用多尺度随机训练,允许网络模型对输入图像大小在尺寸 320—608 进行自由调整,从而提高网络的鲁棒性,一定程度上抑制过拟合现象的发生。为寻求最合适的训练比例,同时观察数据集划分比例对精度的影响,保证不同训练比例下本文算法能保持相对稳定的检测精度,分别使用 50%、60%、70%、80% 的数据集进行训练,剩余数据集进行测试,训练结果如表 2 所示。随

着训练比例从 50% 上升到 80%,检测精度变化不大且维持在一个相对稳定的水平。后续检测结果都以 80% 数据集进行训练、20% 数据集进行测试为基准进行实验。

表 2 不同比例训练结果

Tab. 2 Training results of different scales

Proportion of the training set	mAP
50%	72.36%
60%	72.94%
70%	72.47%
80%	73.28%

3.4 模型评估

在本节中,将提出的网络模型在 GC10-DET 数据集上与 Faster-RCNN、YOLOv2、YOLOv3、YOLOv4 等主流模型进行对比,最后对各项性能进行评估,同时绘制 PR 曲线来评估本文的缺陷检测模型。

表 3 所示为本文所提方法与目前主流网络模型的对比,从表中能够清晰地看到本文所提方法在 Cg 、 Ws 、 In 、 Rp 4 种缺陷中具有最高的检测精度,虽然在 mAP 上不及 Faster-RCNN,但是由于 Faster-RCNN 模型参数量大且检测速度慢,并不适用于硬件部署,也就难以实际应用于带钢表面缺陷检测的工业生产线中。而本文所提方法虽然 mAP 比 Faster-RCNN 低 0.98,但是其检测速度为 70.8 FPS,相比于 Faster-RCNN 高出了 47.2,具备更快的检测速度。在模型大小方面,表 1 中已经提到,本文所提方法的模型参数量和模型复杂度仅为 YOLOv3 的 3.02% 和 4.66%,能够实现较好的轻量化效果。图 11 展示了本文所提模型在 GC10-DET 上对各种缺陷的检测性能。

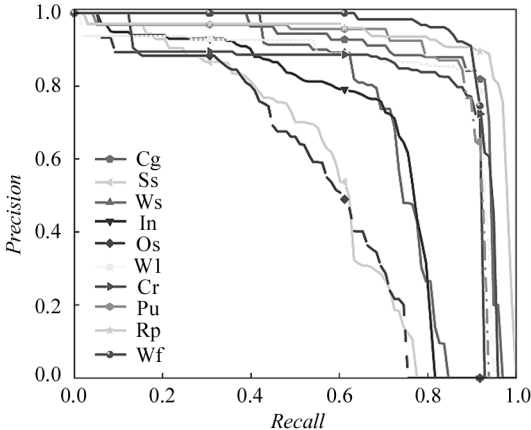


图 11 所提方法的 PR 曲线

Fig. 11 PR curve of the proposed method

表 3 不同算法的对比实验

Tab. 3 Comparative experiments of different algorithms

Type	Faster-RCNN	YOLOv2	YOLOv3	YOLOv3(Res)	YOLOv4	YOLOv4(Res)	Proposed
Pu/%	95.46	95.33	93.44	79.83	88.29	92.45	82.47
Wl/%	88.24	88.70	92.81	77.67	77.65	81.47	75.21
Cg/%	75.65	83.09	77.11	76.18	75.71	82.73	86.49
Ws/%	51.45	52.32	57.80	70.96	56.47	58.18	72.97
Os/%	69.78	58.69	47.76	47.40	61.90	64.92	48.80
Ss/%	50.94	47.44	40.27	47.71	43.60	44.25	48.88
In/%	54.21	51.09	41.59	56.03	64.94	52.30	66.32
Rp/%	87.46	72.09	84.25	82.90	90.09	83.96	92.93
Cr/%	74.95	67.94	78.60	86.08	83.88	76.50	70.82
Wf/%	94.47	91.38	87.71	81.92	73.55	90.85	87.89
mAP /%	74.26	70.81	69.93	70.67	71.61	72.76	73.28
Speed/FPS	23.6	49.9	26.5	41.5	27.1	33.4	70.8

3.5 消融实验

为使模型具有更好的检测效果,本文在轻量化基础上进行了一系列模块的增添,通过消融实验来验证所用方法的有效性。由表 4 可知,原轻量化模型虽然具备了较好的移动端部署能力以及检测速度,但是由于仅靠最后两层进行特征的检测和输出,感受野受限,特征信息利用能力不足,检测精度不够理想。因此,对先验框根据带钢缺陷特点进行重新计算之后,效果提升明显;引入 SPP 模块之后,融合了不同尺度特征图信息,增强了模型的感受野;最后经过融合改进注意力机制的 FPN 模块,给予不同尺度的特征图不同的权重加成,充分利用了特征图信息,从而达到更好的检测效果。图 12 为本文提出的模型在 GC10-DET 带钢缺陷数据集上的检测结果展示。

表 4 消融实验

Tab. 4 Ablation experiment

Methods	mAP	UP
Original+MobileNetv2(CIOU)	67.96%	—
+ Reasonable anchors	69.37%	1.41%
+ SPP	71.21%	1.84%
+ Improved FPN	73.28%	2.07%

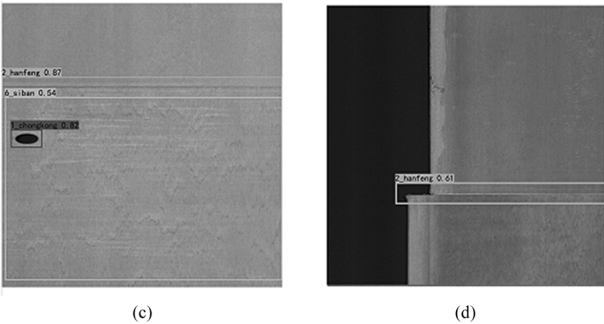
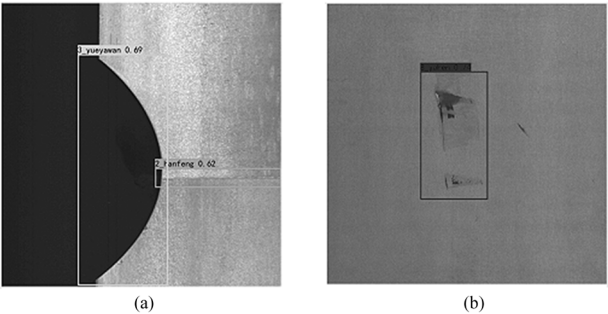


图 12 缺陷检测结果展示:
(a) 月牙形缺口和焊缝缺陷;
(b) 轧坑缺陷; (c) 丝斑、冲孔和焊缝缺陷;
(d) 焊缝缺陷
Fig. 12 Display of defect detection results:
(a) Cg and Wl; (b) Rp; (c) Ss, Pu and Wl; (d) Wl

4 结 论

在本文中,针对现有带钢表面缺陷检测方法存在的算法鲁棒性差、检测速度慢以及移动端部署能力不够理想等问题,提出了一种基于轻量化 YOLOv3 的带钢表面缺陷快速检测方法。FPS 由原网络的 26.5 增长为 70.8,模型参数量也得到了缩减,由此可见改进后的网络较原网络实现了有效的轻量化。在 GC10-DET 数据集上将所提方法与其他主流方法进行对比,结果表明,本文提出的方法在保证检测精度的同时,检测速度明显优于其他模型,对带钢表面缺陷起到了良好的检测效果。对于未来工作,一方面对数据集进行收集扩充,另一方面继续对模型改进,在保证模型轻量化的基础上提升其检测精度。

参考文献:

- [1] FANG X, LUO Q, ZHOU B, et al. Research progress of automated visual surface defect detection for industrial metal planar materials[J]. *Sensors*, 2020, 20(18): 5136.
- [2] HU W, WANG T, WANG Y et al. LE-MSFE-DDNet: a defect detection network based on low-light enhancement and multi-scale feature extraction[J]. *The Visual Computer*, 2022, 38: 3731-3745.
- [3] FANG C H, ZHANG F, DANG W J, et al. Defect identification method for main insulation surface of 10 kV cable based on canny algorithm[J]. *Journal of Optoelectronics • Laser*, 2021, 32(11): 1188-1195.
方春华, 张峰, 党卫军, 等. 基于 Canny 算法的 10 kV 电缆主绝缘表面缺陷识别方法[J]. *光电子 • 激光*, 2021, 32(11): 1188-1195.
- [4] AGHDAM S R, AMID E, IMANI M F. A fast method of steel surface defect detection using decision trees applied to LBP based features[C]//2012 7th IEEE Conference on Industrial Electronics and Applications (ICIEA), July 18-20, 2012, Singapore. New York: IEEE, 2012: 1447-1452.
- [5] LI J F, LI M R. Research on defecting method of light guide plate defect based on machine vision [J]. *Journal of Optoelectronics • Laser*, 2019, 30(3): 256-265.
李俊峰, 李明睿. 基于机器视觉的导光板缺陷检测方法研究[J]. *光电子 • 激光*, 2019, 30(3): 256-265.
- [6] BHATT P M, MALHAN R K, RAJENDRAN P, et al. Image-based surface defect detection using deep learning: A review[J]. *Journal of Computing and Information Science in Engineering*, 2021, 21(4): 1-23.
- [7] GU J, WU Y N, MENG X H. Weld defect detection based on expansion convolution multi-scale fusion[J]. *Journal of Optoelectronics • Laser*, 2022, 33(1): 61-66.
谷静, 吴怡宁, 孟鑫昊. 基于膨胀卷积的多尺度焊缝缺陷检测算法[J]. *光电子 • 激光*, 2022, 33(1): 61-66.
- [8] FANG Y X, GAN P, CHEN L. Improved YOLOv3 algorithm for detection of metal surface defect[J]. *Mechanical Science and Technology for Aerospace Engineering*, 2020, 39(9): 1390-1394.
方叶祥, 甘平, 陈俐. 金属表面缺陷检测的改进 YOLOv3 算法研究[J]. *机械科学与技术*, 2020, 39(9): 1390-1394.
- [9] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 4510-4520.
- [10] TAN M, PANG R, LE Q V. Efficientdet: Scalable and efficient object detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020, Seattle, WA, USA. New York: IEEE, 2020: 10781-10790.
- [11] WOO S, PARK J, LEE J Y, et al. CBAM: Convolutional block attention module [C]//European Conference on Computer Vision (ECCV), September 8-14, 2018, Munich, Germany. Berlin: Springer, 2018: 3-19.
- [12] REDMON J, FARHADI A. Yolov3: An incremental improvement[EB/OL]. (2018-04-08) [2022-06-16]. <https://arxiv.org/abs/1804.02767>.
- [13] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.
- [14] HOWARD A G, ZHU M, CHEN B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications [EB/OL]. (2017-04-17) [2022-06-16]. <https://arxiv.org/abs/1704.04861>.
- [15] SHAO M, ZHANG W, LI Y, et al. Enhanced feature pyramidal network for object detection[J]. *Journal of Electronic Imaging*, 2022, 31(1): 013030.
- [16] 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.
- [17] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//AAAI Conference on Artificial Intelligence, February 7-12, 2020, New York, NY, USA. Palo Alto: AAAI, 2020, 34(7): 12993-13000.

作者简介:

刘国华 (1969—), 男, 硕士, 教授, 硕士生导师, 主要从事计算机视觉与图像处理方面的研究。