计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(4):248−254 [doi: 10.15888/j.cnki.csa.009075]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 Transformer 和注意力机制的角钢塔螺栓缺陷检测模型①

程智余[1], 张金锋[1], 孙丙宇[2]

[1](国网安徽省电力有限公司, 合肥 230022)
[2](中国科学院 合肥智能机械研究所, 合肥 230031)
通信作者: 张金锋, E-mail: 5952544@qq.com

摘　要: 螺帽缺失、螺栓缺失是角钢塔建设阶段常见的结构缺陷, 但由于特征区分度低现有目标检测算法对螺栓缺陷检出率较低. 针对这个问题, 首先基于 Transformer 对卷积特征进行特征编码提出了全局信息提取算子, 其次通过通道注意力机制自适应组合候选检测框多尺度缩放后引入的局部背景信息, 最后基于图像分割与背景融合对螺栓缺陷样本进行数据扩增. 消融实验表明上述策略均能有效提升螺栓缺陷检测效果且相互不排斥, 与其他典型算法对比验证了本文算法的先进性.

关键词: 角钢塔螺栓; 缺陷检测; Transformer 编码器; 通道注意力机制

引用格式: 程智余,张金锋,孙丙宇.基于 Transformer 和注意力机制的角钢塔螺栓缺陷检测模型.计算机系统应用,2023,32(4):248–254. http://www.c-s-a.org.cn/1003-3254/9075.html

Defect Detection Model of Angle Steel Tower Bolt Based on Transformer and Attention Mechanism

CHENG Zhi-Yu[1], ZHANG Jin-Feng[1], SUN Bing-Yu[2]

[1](State Grid Anhui Electric Power Co. Ltd., Hefei 230022, China)
[2](Institute of Intelligent Machine, Chinese Academy of Sciences, Hefei 230031, China)

Abstract: The missing of nuts and bolts is a common structural defect in the construction stage of angle steel towers, but the detection rate of bolt defects by existing object detection algorithms is low due to low feature discrimination. In order to solve this problem, a global information extraction operator is proposed based on Transformer to encode convolutional features. Secondly, the local background information introduced after the multi-scale scaling of the candidate detection frame is adaptively combined through the channel attention mechanism. Finally, the bolt defect samples are amplified based on image segmentation and background fusion. The ablation experiments show that the above strategies can effectively improve the detection effect of bolt defects and do not exclude each other. Compared with other typical algorithms, this algorithm has been proven to be advanced.

Key words: angle steel tower bolt; defect detection; Transformer encoder; channel attention mechanism

随着我国工业生产用电的需求的提高, 近几年国家大力推动电力传输网络的建设, 其中架设角钢塔输送电由于经济高效的优势得到广泛推广. 然而, 在角钢塔建设过程中由于装配环境恶劣、作业流程繁复枯燥, 工人在角钢塔装配过程中常常会出现螺栓漏装、螺帽漏拧的问题, 这些情况会导致角钢塔结构整体刚度减弱, 进而对角钢塔的受力产生极大的影响[1], 因此, 对角钢塔的螺栓安装状况进行复查是必不可少的装配质量保证措施. 近些年一些无人机和攀爬机器人陆续应用到电力巡检当中[2,3], 巡检工人操控这些设备, 通过搭载

① 基金项目: 国网安徽省电力有限公司科技项目 (52120019007G); 安徽省能源互联网基金 (2008085UD03)
收稿时间: 2022-08-19; 修改时间: 2022-09-27, 2022-10-12; 采用时间: 2022-11-30; csa 在线出版时间: 2023-03-01
CNKI 网络首发时间: 2023-03-02

的摄像头回传的视频画面进行螺栓缺陷检测, 这种方式减轻了人工攀爬的工作负担, 但仍然需要人工从视频上检查螺栓缺陷, 自动化程度低.

伴随着深度学习的发展, 基于视觉的自动化检测算法在工业上也得到广泛应用, 甚至超过了人眼的检测水平. Zhao 等人[4] 使用了多个卷积神经网络模块提取图像特征进行电力传输线路绝缘检测. Tao 等人[5] 则提出了基于 CNN 级联结构的绝缘线路检测. 然而, 由于角钢塔螺栓目标较小, 以及螺栓位置复杂多变, 关于螺栓的缺陷检测研究甚少. 除此之外, 光照和拍摄姿态的变化, 也增加了检测的难度. 当前基于深度学习的目标检测算法在计算机视觉领域取得了重大突破, 主要分为单阶段和双阶段两类. 双阶段的主要代表是区域卷积神经网络 (region convolutional neural network, R-CNN)[6] 系列, 而单阶段主要是单阶段检测器 (single-shot detector, SSD)[7] 和 YOLO (you only look once)[8].

最近, 基于自注意力机制的 Transformer[9] 在自然语言处理领域获得了巨大的成功, 部分学者也将其应用计算机领域展开研究[10], 使得目标检测算法又出现新的发展. Transformer 比 CNN 和循环神经网络 (recurrent neural network, RNN) 具有更多的优势, 与 CNN 的局部感受野相比 Transformer 具有全局感受野, 与 RNN 相比具有更强的捕捉长期上下文依赖, 具有并行化计算的优势, 这与顺序输入的 RNN 相比有着更大工程部署可行性. 因此, Transformer 引起目标检测研究领域学者的广泛关注, Detection Transformer (DETR) 是第一个将 Transformer 用于目标检测的模型, 并且在 COCO 数据集上获得了竞争性结果[11], 然而 DETR 收敛性较差并且对于小目标检测效果不佳. Vision Transformer (ViT)[9] 是第一个纯 Transformer 在大规模数据集上取得先进结果的模型. ViT 直接将输入的图片切成 16×16 的小块, 然后通过线性投影 (linear projection) 后输入到 Transformer 的编码器 (encoder), 最后通过多层感知机 (multi-layer perceptron, MLP) 预测类别. ViT 的成功应用引起了诸多后续研究. ViT-FRCNN[12] 将 Faster R-CNN 的骨干网络换成 ViT, 最后通过区域建议网络 (region proposal network, RPN) 预测结果.

在本文针对角钢塔中螺帽缺失误检测率高、螺栓缺失漏检测率高的问题展开研究, 提出了一种纯视觉的螺栓缺陷检测方式. 本文首先参考 ViT-FRCNN 设计了基于 Transformer 的全局信息提取算子, 并将原始 ViT-FRCNN 输入部分的图片块嵌入替换成一个 CNN 骨干网络, 这避免了 Transformer 编码空间信息提取不足的问题. 同时本文对区域推荐的网络给出的推荐检测框进行多尺度变换引入局部信息, 使用通道注意力机制对有效的局部信息进行重点关注. 为弥补缺陷数据收集不足的问题, 本文还基于图像分割与背景融合提出了缺陷样本扩增策略. 最后, 本文通过消融实验和不同算法对比实验, 证明了本文算法对角钢塔螺帽缺失、螺栓缺失检测的先进性.

1 基于 Transformer 和注意力机制的角钢塔螺栓缺陷检测模型

螺帽缺失、螺栓缺失是角钢塔中常见的两种螺栓缺陷, 典型示例如图 1 所示, 其与普通的目标检测具有自身的特殊性. 螺帽缺失会遗留下螺柱, 但由于螺柱、螺帽的型材相近, 并且螺帽是套在螺栓上面的, 因此直接使用常规的目标检测算法容易将正常套有螺帽的螺栓误识别成螺帽缺失. 螺栓缺失后只会留下螺孔, 但螺孔后面的背景信息复杂多样且只会带来干扰信息, 真正利于螺孔检测的信息只是螺孔周围的一圈边缘梯度信息, 常规的目标检测算法会存在严重的漏检.

(a) 正常螺栓 (b) 螺帽缺失 (c) 螺栓缺失

图 1 正常螺栓与螺栓缺陷示例

针对上述问题, 本文基于 Transformer 和注意力机制提出一套新的角钢塔螺栓缺陷检测模型, 综合图像的全局与局部信息降低螺帽缺失误识别和螺栓缺失漏检测的发生. 总体模型流程框架如图 2 所示, 输入图片先经过卷积神经网络进行特征提取, 然后将卷积特征分成 9 块输 Transformer 编码器中进行全局信息提取, 将编码后的特征按照分割顺序重新拼接, 基于 Faster R-CNN 中的区域推荐网络生成候选区域, 对候选区域使用局部信息提取算子进行特征提取. 将引入局部上下文信息的候选区域特征经全连接层处理后分别进行回归和 Softmax 操作, 最终得到螺栓缺陷的精确检测框坐标与缺陷类别.

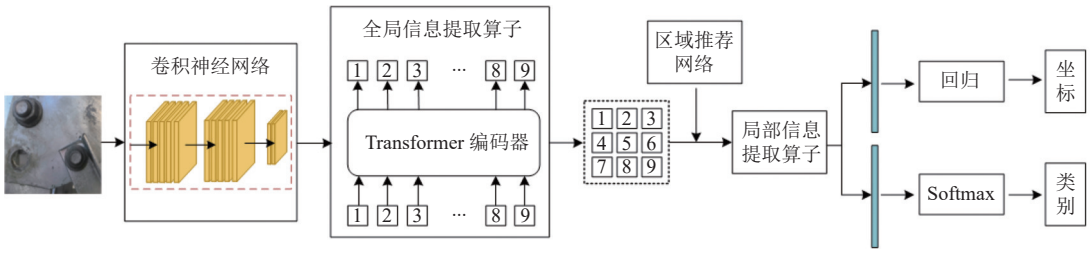


图 2 算法框架示意图

1.1 基于 Transformer 的全局信息提取算子

考虑到螺栓、螺孔体积很小, 只占图像中较小的区域, 同时螺栓、螺孔通常出现在角钢塔的斜材的固定位置, 因此本文算法首先参考 ViT-FRCNN 中的 Transformer 思想进行全局信息提取, 期望基于全局信息引导模型对螺栓、螺孔可能出现的重点区域进行关注.

Transformer 利用注意力机制建立起序列之间的远距离依赖关系, 相比卷积神经网络具有更强的建模能力. 但其同样丢失了卷积神经网络平移不变性等仿生学特性, 并且如 ViT 论文所阐述 Transformer 只有在大数据集上才能取得优于卷积神经网络的效果. 针对 Transformer 的这一不足, 本文并不是如 ViT-FRCNN 那样直接对图片进行分块并送入 Transformer 编码器中, 而是先利用卷积神经网络进行特征提取后切分输入编码器.

Transformer 的编码器结构如图 3 所示, 主要是由多头自注意力、多层感知机、残差连接[13] 以及层归一化[14] 组成的. 最核心的自注意力机制表示为:

$$Attention(Q,K,V)=Softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{1}$$

其中, Q是查询 (query) 矩阵, K是键 (key) 矩阵, V是值 (value) 矩阵, d_k是V的维度. QK^{T}计算出不同输入矩阵间的注意力分数, 缩放因子$1/\sqrt{d_k}$主要用来提高稳定性, 然后$Softmax$函数将注意力分数转化为概率. 最后, 再乘上V获得权重矩阵.

为了增强 self-attention 的特征提取能力, 将多个 self-attention 拼接成多头注意力, 具体可以写为:

$$MSA(Q,K,V)=Concat(head_1,\cdots,head_n)W^0 \tag{2}$$

$$head_i=Attention(QW_i^Q,KW_i^K,VW_i^V) \tag{3}$$

其中, W表示权重矩阵, 多头的数量默认设置为 6.

除了自注意力机制和多头自注意力之外, 层归一化主要是用来稳定训练和加速收敛, 残差连接则是提高信息流. 多层感知机由两个线性层和一个高斯误差线性单元[15] 激活函数组成. 最后, 通过反复堆叠编码器不断地提取全局上下文依赖和聚合特征信息.

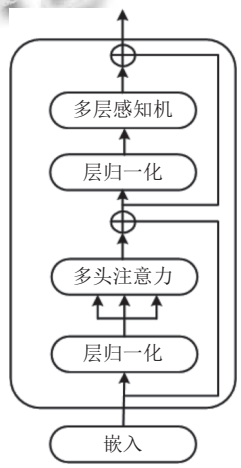


图 3 Transformer 编码器的结构

1.2 基于注意力机制的多尺度局部信息提取算子

当前检测算法的检测效果受限于区域推荐网络的性能, 对检测目标的分类与检测框回归只是单纯只利用建议区域内的特征. 本文参考文献 [16] 提出了基于注意力机制提出了多尺度局部信息提取算子, 突破了原始区域推荐网络对检测结果的限制, 引入了建议区域周围的局部信息, 并通过通道注意力机制赋予不同尺度的特征以不同权重, 局部信息提取算子的具体网络结构如图 4 所示.

基于区域推荐网络产生建议区域, 每个建议区域的检测框的长宽首先使用 3 个预先设置的缩放因子进行缩放, 使得原始预测的候选检测框能够在不同尺度上包含局部背景信息. 紧接着对不同尺度的候选区域特征使用 ROIPooling 转化为相同的尺度, 进一步经过 L2 归一化后沿通道拼接起来产生一个融合的特征. 考虑到不同尺度的局部信息对最终的缺陷检测会有不同程度的贡献度, 例如在原始候选检测框预测过小时候

就需要赋予大尺度因子的特征更高的权重, 因此基于通道注意力模块对融合后的特征的不同通道上赋予不同的权重, 将最终处理后的特征用于最终的目标边界框回归和类别预测.

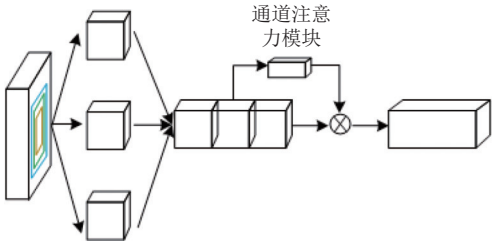


图 4 局部信息提取算子

2 模型训练

2.1 基于图像分割与背景融合的缺陷样本扩增策略

由于深度神经网络是一种极易过拟合的算法, 但螺栓缺失、螺帽缺失样本在真实环境出现的概率并不高, 因此本节提出了一种缺陷样本扩增策略. 预先标注一些缺陷区域的图像分割样本, 螺帽缺失可直接将留下的螺柱区域标注出来, 而由于螺栓缺失留下的是一个空洞, 标注时候需要将空洞外扩后进行标注. 基于标注好的数据对 U-net 分割网络进行训练, 然后如图 5 所示进行缺陷区域分割, 为保证后期扩增的缺陷样本的可靠性需要进行人工筛选.

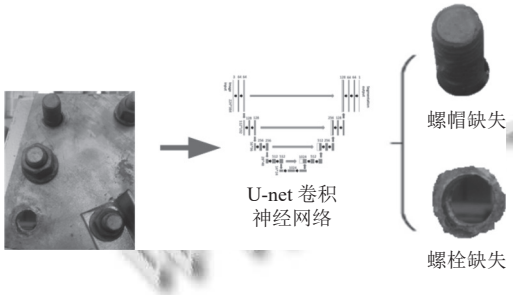


图 5 基于 U-net 的缺陷区域分割

仿射变换是一种二维坐标(x,y)到二维坐标(u,v)的线性变换, 其表达式形式如下:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} M & t \\ 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{4}$$

其中, 分块矩阵M能够控制图像的尺度变换、旋转变换和反转变换, 而矩阵t则是可以控制图像的平移变换, 对分割出的缺陷区域如图 6 所示进行仿射变换.

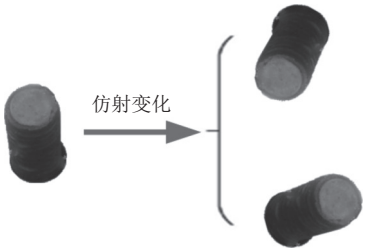


图 6 缺陷区域的仿射变换

如图 7 所示, 将仿射变换后的缺陷区域与不同的背景进行融合即可生成新的缺陷样本, 具体融合策略如式 (5) 表示:

$$I_{\text{new}}(x,y) = \begin{cases} \lambda \cdot I_{fg}(x,y) + (1-\lambda) \cdot I_{bg}(x,y), \text{ if } I_{fg}(x,y) \neq 0 \\ I_{bg}(x,y), \text{ 其他} \end{cases} \tag{5}$$

其中, $I_{fg}(x,y)$表示前景缺陷部件图像, $I_{bg}(x,y)$表示背景图像. 参数λ是图像融合参数, 其取值范围为 [0, 1], 该参数是为了使缺陷部件图像与新背景图像的边缘过渡平滑, 使他们能够更好地融合在一起.

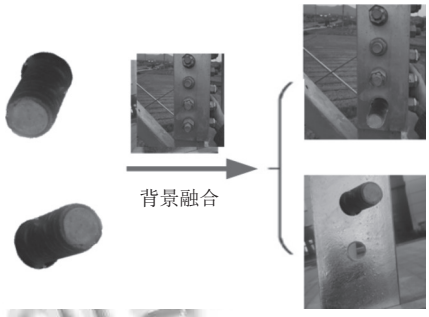


图 7 背景融合效果

随机设置仿射变换参数将 U-net 分割出来的螺帽缺失、螺栓缺失缺陷进行变换以扩增缺陷形态多样性, 将变换后的螺栓缺陷在四周进行补 0 来将分割出的缺陷图像扩充到与背景图像同样尺寸大小. 为保证融合的数据更加符合真实情况, 避免螺栓缺陷与非角钢塔区域进行融合. 对背景图片预先粗略标注好角钢塔可能的融合区域, 需要控制螺栓缺陷总在补 0 后图像的角钢塔区域. 通过上述的前景背景融合公式将补 0 后的螺栓缺陷图像与正常的背景图像进行融合, 最终获得扩增的螺栓缺陷样本图像.

2.2 实验配置

实验设备为 Intel i9-10900K 处理器, 32 GB 内存的高性能服务器, 配置有 Nvidia GeForce GTX 1080Ti 显卡, 检测模型采用 PyTorch 深度学习框架搭建.

实验采用随机梯度下降优化器, 并配置了 0.9 的动量, 权重衰减设置为 0.000 1, 初始学习率设这位 0.08, 批量大小 (batch size) 为 64. 输入图片尺寸设置为 512×512, 对于特征提取的卷积神经网络采用预训练初始化的 ResNet50 网络, 在训练过程中采用了水平翻转、光照增强、mosaic 等数据增强策略. 经过收集与上节的数据扩增策略总共收集到了 5 926 张图片, 随机选择其中的 3 000 张图片作为训练集, 剩下的 2 926 张图片作为测试集.

3 实验分析

3.1 评价指标

采用常用的评估指标对模型进行定量评估, 包括准确率 (precision, P)、召回率 (recall, R), 精度均值 (mean average precision, mAP). 其中, 准确率和查全率的计算公式分别如下:

$$P = \frac{TP}{TP + FP} \tag{6}$$

$$R = \frac{TP}{TP + FN} \tag{7}$$

其中, TP是真正例, FN是假反例, FP是假正例. mAP 是一种广泛用于目标检测的指标, 主要用于评估预测框的准确性.

3.2 消融实验

为了更充分地分析本文所提出的全局信息提取算子、局部信息提取算子、缺陷样本数据扩增对最终角钢塔缺陷检测的贡献度, 在本节将这些算法技巧分别记为 1、2、3, 以 Faster R-CNN 算法为基底进行消融实验. 为保证实验结果的可靠性, 消融实验在相同配置的服务器上使用相同的数据进行训练与测试, 具体实验结果如表 1 所示.

表 1 中的第 1 行“基底”指使用原始的 Faster R-CNN 进行训练, 第 2 行“1”指在原始的 Faster R-CNN 的卷积特征提取后面添加了基于 Transformer 的全局信息提取算子, 第 3 行“2”指代在原始的 Faster R-CNN 的区域推荐网络后面添加了基于注意力机制的多尺度局部信息提取算子, 第 4 行“1+2”代表对原始的 Faster R-CNN 同时添加了全局信息提取算子与局部信息提取算子, 第 5 行“1+2+3”相对第 4 行使用了基于图像分割与背景融合的缺陷样本扩增策略对训练数据进行了扩增.

表 1 消融实验结果 (%)

算法	螺栓缺失			螺帽缺失		
	准确率	召回率	mAP	准确率	召回率	mAP
Faster R-CNN	83.8	88.4	79.7	84.6	79.1	75.2
1	84.2	91.2	81.9	84.8	81.8	77.1
2	85.6	89.7	81.4	85.2	80.8	77.6
1+2	86.3	92.4	82.3	86.7	82.3	78.3
1+2+3	86.7	93.1	82.7	87.3	82.5	78.6

将第 2 行、第 3 行实验结果与第 1 行原始的 Faster R-CNN 算法对比, 可以看出本文提出的全局信息提取算子、局部信息提取算子对角钢塔螺栓缺陷检测均有一定贡献, 其中全局信息提取算子通过全局信息的引导明显提升了螺栓缺失、螺帽缺失的召回率, 而局部信息提取中的注意力机制与局部信息利用提升了螺帽缺失与正常螺栓的区分度, 也增加了螺栓缺失中螺孔边缘信息的利用. 从第 4 行实验结果可以看出, 全局信息提取算子、局部信息提取算子两者并不矛盾, 两者的综合使用能取得更佳的检测效果. 从第 5 行实验结果可以看出, 本文基于图像分割与背景融合提出的缺陷样本扩增策略能一定程度上缓解训练数据不足缺陷, 提升检测效果.

3.3 与其他算法对比

为了进一步验证本文针对角钢塔螺栓缺陷检测常见所提出算法的有效性, 在本节选择了 3 种经典目标检测模型与其他进行对比, 分别为 SSD[7]、ViT-FRCNN 和 YOLOv3[17], 选择相同的实验平台与训练数据进行模型训练来确保对比的公平性.

(1) 定量对比

根据测试结果计算上述评估指标后, 得到如表 2 所示的实验结果.

表 2 不同算法定量对比 (%)

算法	螺栓缺失			螺帽缺失		
	准确率	召回率	mAP	准确率	召回率	mAP
SSD	82.3	87.2	77.9	82.8	79.1	74.5
ViT-FRCNN	83.6	91.8	80.4	84.2	81.0	76.1
YOLOv3	85.5	92.1	81.5	85.1	81.6	76.9
本文算法	86.7	93.1	82.7	87.3	82.5	78.6

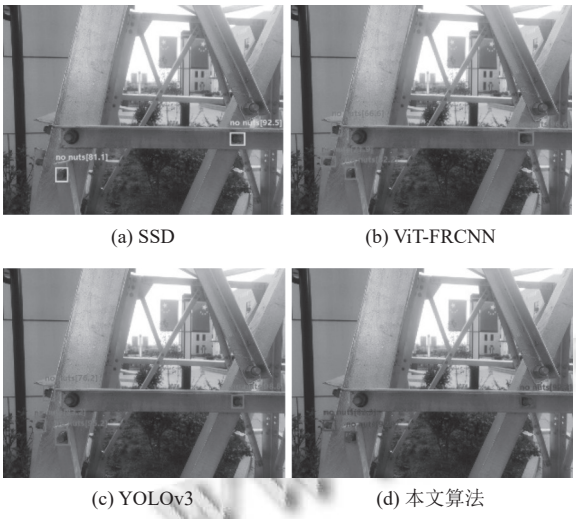
通过对比实验结果可以发现, SSD 对螺帽缺失、螺栓缺失检测的准确率、召回率、mAP 均是最低的, 这主要是由于 SSD 算法只是简单地将不同层次特征提取出来用于检测并没有做进一步加工, 而 ViT-FRCNN 基于 Transformer 对原始图片进行编码提升检

测时候全局信息的利用率, 但其只是在图片级别进行编码丧失了像素之间的空间结构信息, YOLOv3 算法通过特征融合、IOU 重定义、损失函数改进等策略取得了较好的效果, 在 3 种公开算法中取得了最好检测结果.

本文的算法通过 Transformer 对图片的全局信息进行提取, 基于注意力机制对多尺度局部特征进行融合, 最终无论是螺帽缺失还是螺栓缺失均取得了最佳的检测效果, mAP 值相对次好地 YOLOv3 算法分别提升了 1.2% 和 1.7%, 这验证了本文针对角钢塔螺栓缺陷检测场景设计进行的算法改进是合理而有效的.

(2) 定性对比

为了对 4 种算法对角钢塔螺栓缺陷检测效果有个直观对比, 在图 8 和图 9 中分别展示了 4 种算法对螺帽缺失、螺栓缺失检测效果定性对比, 其中左上角是 SSD 算法的检测结果, 右上角是 ViT-FRCNN 算法的检测结果, 左下角是 YOLOv3 算法的检测结果, 右下角是本文算法的检测结果.



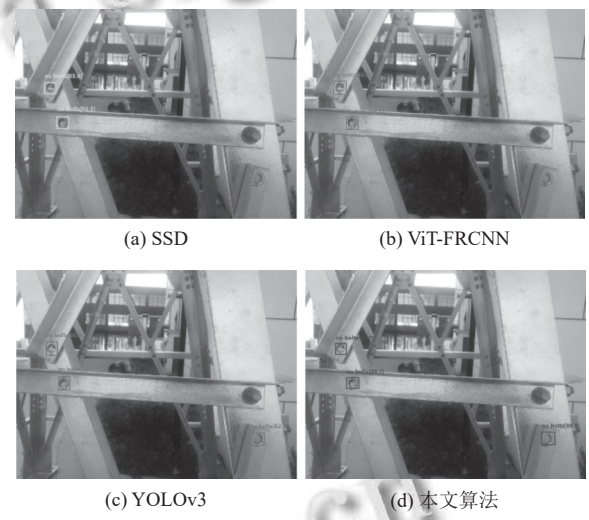
(a) SSD (b) ViT-FRCNN

(c) YOLOv3 (d) 本文算法

图 8 螺帽缺失检测效果对比

图 8 展示了 4 种算法对螺帽缺失的检测效果, 可以看到 SSD 算法正面螺帽缺失检测效果较好, 但是侧面的螺帽缺失存在漏检. ViT-FRCNN 算法与 YOLOv3 算法相对 SSD 算法正面检测检测的置信度更高, 两者也均可以检测到侧面螺帽缺失, 但也均将正常螺栓误检测为螺帽缺失了. 本文提出的算法取得的检测效果最好, 无论正面还是侧面的螺帽缺失均能正确检测出来, 侧面的正常螺栓也没有误检测, 同时检测的置信度值也较高, 这主要归功于本文的全局与局部信息提取算子, 两者均极大地提升了图片中螺栓区域的关注度, 在避免误检测的发生同时提升了检测的置信度.

图 9 展示了螺栓缺失场缺陷的检测效果, 可以看到图片中共有 3 处螺栓缺失, 其中 SSD 与 ViT-FRCNN 算法均只识别到了 2 处缺陷, YOLOv3 和本文算法对 3 处螺栓缺失均正确识别, 这主要是地处螺栓缺失缺陷由于拍摄角度问题被斜材有部分遮挡导致边缘不是一个完整的圆. 在 4 种算法中本文的算法检出率最高同时检测置信度也较高, 这体现全局信息对缺陷检测的引导作用, 局部信息又进一步提升了算法检测效果.



(a) SSD (b) ViT-FRCNN

(c) YOLOv3 (d) 本文算法

图 9 螺栓缺失检测效果对比

4 结论与展望

本文首先对角钢塔巡检作业中的螺帽缺失、螺栓缺失的检测难点进行分析, 然后基于 Transformer 对卷积神经网络提取的特征进行编码提出了全局信息提取算子, 其有效提升了模型对螺栓、螺孔区域的关注度. 通过对候选检测框进行多尺度缩放进一步引入了局部背景信息, 并通过空间注意力机制对多尺度局部信息进行自适应组合. 针对真实场景螺栓缺陷数据收集困难的问题, 本文基于图像分割与背景融合提出了一种缺陷样本数据扩增策略. 在最后设计了消融实验验证了本文提出的全局信息提取算子、局部信息提取算子、数据扩增策略的有效性, 并通过与其他算法对比进一步验证本文算法在角钢塔螺栓缺陷检测场景的优势.

参考文献

1 王俊超, 鞠彦忠, 王德弘, 等. 关键位置螺栓脱落对角钢塔抗风性能影响分析. 东北电力大学学报, 2019, 39(3): 79–84.

2 冯积家. 基于 5G 无人机在电力输电线路自动巡线的实现与研究. 电力设备管理, 2020, (9): 203–204.

3 陈亮, 陈定君. 电力输电线路绝缘子裂缝智能识别机器人设计. 自动化与仪器仪表, 2021, (5): 197–201.

4 Zhao ZB, Xu GZ, Qi YC, *et al*. Multi-patch deep features for power line insulator status classification from aerial images. Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN). Vancouver: IEEE, 2016. 3187–3194.

5 Tao X, Zhang DP, Wang ZH, *et al*. Detection of power line insulator defects using aerial images analyzed with convolutional neural networks. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 50(4): 1486–1498. [doi: 10.1109/TSMC.2018.2871750]

6 Girshick R, Donahue J, Darrell T, *et al*. Region-based convolutional networks for accurate object detection and segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142–158. [doi: 10.1109/TPAMI.2015.2437384]

7 Liu W, Anguelov D, Erhan D, *et al*. SSD: Single shot multibox detector. Proceedings of the 14th European Conference on Computer Vision. Cham: Springer, 2016. 21–37.

8 Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788.

9 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

10 Dosovitskiy A, Beyer L, Kolesnikov A, *et al*. An image is worth 16x16 words: Transformers for image recognition at scale. Proceedings of the 9th International Conference on Learning Representations. OpenReview.net, 2021. 1–21.

11 Carion N, Massa F, Synnaeve G, *et al*. End-to-end object detection with transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229.

12 Beal J, Kim E, Tzeng E, *et al*. Toward transformer-based object detection. arXiv:2012.09958, 2020.

13 Wang F, Jiang MQ, Qian C, *et al*. Residual attention network for image classification. Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu: IEEE, 2017. 6450–6458.

14 Ba JL, Kiros JR, Hinton GE. Layer normalization. arXiv:1607.06450, 2016.

15 Hendrycks D, Gimpel K. Gaussian error linear units (GELUs). arXiv:1606.08415, 2016.

16 Li JN, Wei YC, Liang XD, *et al*. Attentive contexts for object detection. IEEE Transactions on Multimedia, 2017, 19(5): 944–954. [doi: 10.1109/TMM.2016.2642789]

17 Redmon J, Farhadi A. YOLOv3: An incremental improvement. arXiv:1804.02767, 2018.

(校对责编: 牛欣悦)