ORIGINAL RESEARCH

AMCD: an accurate deep learning-based metallic corrosion detector for MAV-based real-time visual inspection

Leijian Yu[1] · Erfu Yang[1] · Cai Luo[2] · Peng Ren[2]

Abstract

Corrosion has been concerned as a serious safety issue for metallic facilities. Visual inspection carried out by an engineer is expensive, subjective and time-consuming. Micro Aerial Vehicles (MAVs) equipped with detection algorithms have the potential to perform safer and much more efficient visual inspection tasks than engineers. Towards corrosion detection algorithms, convolution neural networks (CNNs) have enabled the power for high accuracy metallic corrosion detection. However, these detectors are restricted by MAVs on-board capabilities. In this study, based on You Only Look Once v3-tiny (Yolov3-tiny), an accurate deep learning-based metallic corrosion detector (AMCD) is proposed for MAVs on-board metallic corrosion detection. Specifically, a backbone with depthwise separable convolution (DSConv) layers is designed to realise efficient corrosion detection. The convolutional block attention module (CBAM), three-scale object detection and focal loss are incorporated to improve the detection accuracy. Moreover, the spatial pyramid pooling (SPP) module is improved to fuse local features for further improvement of detection accuracy. A field inspection image dataset labelled with four types of corrosions (the nubby corrosion, bar corrosion, exfoliation and fastener corrosion) is utilised for training and testing the AMCD. Test results show that the AMCD achieves 84.96% mean average precision (mAP), which outperforms other state-of-the-art detectors. Meanwhile, 20.18 frames per second (FPS) is achieved leveraging NVIDIA Jetson TX2, the most popular MAVs on-board computer, and the model size is only 6.1 MB.

Keywords Corrosion detection · High accuracy · MAVs · Deep learning

1 Introduction

Corrosion is a major threat to metallic facilities for industries, as it gradually reduces the strength of metallic assets (Gomes et al. 2013). Additionally, failures of these infrastructures caused by corrosions will lead to unacceptable safety concerns for the public and environmental damage. Thus, regular inspection and maintenance of metallic structures are required (Tscheliesnig et al. 2016).

Visual inspection is one of the most basic and reliable inspection techniques (Moosavi 2017). These assets are traditionally inspected by experienced human engineers that mainly rely on their naked eyes. However, lots of facilities are set in the harsh or cluttered environment where it is hard to reach for humans. Therefore, liberating human engineers from dangerous, expensive and time-consuming tasks becomes urgent (Agnisarman et al. 2019).

Currently, with the development of autonomous Robotics and Autonomous Systems (RAS), an increasing number of companies choose to exploit robots for smart inspection. Among them, Micro Aerial Vehicles (MAVs) have gained great interest due to their flexibility and manoeuvrability. MAVs inspection of facilities is able to increase the diagnostic speed and reduce the costs associated with the inspection procedure (Chen et al. 2019). Furthermore, thanks to advanced image processing and MAVs technologies, there is an opportunity to deploy these technologies for automated and efficient inspection of facilities in industries (Hoskere et al. 2018). To develop MAV-based autonomous visual inspection systems, high-accuracy corrosion detectors based

✉ Erfu Yang
erfu.yang@strath.ac.uk

[1] Department of Design, Manufacturing and Engineering Management, University of Strathclyde, Glasgow G1 1XJ, UK

[2] College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China

on advanced computer vision techniques will be the primary concern (Atha and Jahanshahi 2018).

Over the last couple of years, a variety of algorithms for corrosion detection have been proposed. Among them, texture and colour analyses by a filter-based approach or a statistical model have gained great interest. The colour wavelet filter bank is one of the most popular techniques, which detects corrosion through filtering texture and colour features. However, when the optimal features are not identified, the detection accuracy will decrease heavily (Chu and Thuerey 2017).

Recently, Convolution Neural Networks (CNNs) have been proved to surpass humans on the ImageNet classification (He et al. 2015a). According to our investigation, an increasing number of researchers have adopted CNNs to assist their research, such as morbidity identification (Kumar et al. 2021), SAR image classification (Gao et al. 2017a), vehicle detection (Chen et al. 2020), wind turbine blade structural state evaluation (Sarkar and Gunturi 2020) and bridge crack detection (Xu et al. 2019). These results suggest that CNNs could also be utilised to achieve high accuracy corrosion detection. Unlike previous approaches, CNNs do not need prior designed low-level features, which are not robust enough for computer vision tasks. For CNN-based computer vision tasks, features are determined inherently by CNNs and the training dataset. The results in Guindel et al. (2017) indicated that CNNs are robust enough to detect or classify objects with different scales, orientations and illuminations. Thus, an opportunity has emerged for CNN-based detectors to achieve much more accurate corrosion detection than traditional approaches.

Although several existing works have shown accurate corrosion detection with CNNs, the high computational cost of image processing still poses a challenge in adopting these methods onto a low-cost and low-power computing platform (the MAVs on-board computer). These researches only focus on image processing without consideration of limitations of the MAV platform. In this paper, an accurate deep learning-based corrosion detector is proposed for real-time implementation on MAVs on-board platforms. To our best knowledge, this is the first deep learning-based MAV on-board real-time corrosion detector. The main contributions of this paper are as follows:

1. A novel accurate deep learning-based metallic corrosion detector (AMCD) is proposed, which is able to achieve high accuracy detection with acceptable frames per second (FPS) on the off-the-shelf commercial MAVs on-board computer.
2. Depthwise separable convolution (DSConv) layers are applied to corrosion detection for the first time to reduce model parameters significantly and preserve detection accuracy.
3. An improved Spatial Pyramid Pooling (SPP) module is presented to fuse local features for further improving the detection accuracy. Moreover, the convolutional block attention module (CBAM), three-scale detection and focal loss are adopted in the corrosion detector for further enhancement of detector performance.
4. Comprehensive validation experiments, analysation and comparisons are performed to evaluate the effectiveness of the AMCD in both corrosion detection accuracy and efficiency.

The rest of this article is organised as follows. In Sect. 2, a comprehensive overview of the related literature on corrosion detection is presented. Details of the proposed detector are given in Sect. 3. Section 4 shows the experimental environment and results. Section 5 concludes the whole article.

2 Related works

2.1 Low-level feature-based corrosion detectors

Apparently, feature extraction plays a crucial role in corrosion detection. Colour, as one of the most basic and popular features, is widely used for computer vision tasks. Bonnín-Pascual and Ortiz (2010) trained a classifier to classify corrosions which used a code-word dictionary consisting of the stacked histogram for red, green and blue colour channels. Utilising colour information for corrosion detection was further investigated in Bonnin-Pascual and Ortiz (2014). As Hue-Saturation-Value (HSV) values of corrosion areas are confined in the Hue-Saturation (HS) plane, they utilised a classifier that works over HSV space to recognise corrosions. Shapes and sizes of corrosions were applied to detect the pitting corrosion in Pereira et al. (2012). The texture analysis for corrosion detection was proposed in Hoang and Tran (2019). In their theory, based on image colour, gray-level co-occurrence matrix (GLCM) and gray-level run lengths (GLRL), 78 features were extracted from the corrosion area. After that, a decision boundary for classifying corrosion images was constructed by the support vector machine (SVM). In Hoang (2020), the texture analysis was utilised for pitting corrosion detection. Statistical measurements of colour channels, GLCM and local binary pattern were computed to characterise properties of the metal surface, and 93 texture features were obtained. The SVM was then employed to detect the pitting corrosion. These traditional approaches require previous knowledge about corrosions and their optimal features. However, how to determine optimal

features of corrosions is still challenging (Khaire and Dhanalakshmi 2019).

2.2 Deep learning-based corrosion detectors

For CNN-based approaches, it is simple to determine the features autonomously, which is able to avoid the requirement of prior information. Several studies on high-accuracy corrosion detection with CNNs have already been proposed recently. Atha and Jahanshahi (2018) finetuned a CNN network to classify and identify the corrosion position through the sliding window technique. Based on corrosion levels, Du et al. (2018) proposed a two parallel CNN architecture to classify corrosions. Apart from aforementioned approaches, there are also some other works that adopted CNN-based object detection approaches to locate corrosions. Faster RCNN was trained by 1737 images to detect the steel corrosion and bolt corrosion in Cha et al. (2018). Li et al. (2018) modified the You Only Look Once (Yolo) architecture to all convolutional layers to detect corrosions of flat steel. These works achieved satisfying corrosion classification or detection results in their scenarios. However, they only focus on image processing without consideration of MAV-based applications. Their networks contain a large number of standard convolutions, resulting in a large model for the entire network. For this reason, these approaches cannot be applied to MAVs on-board platforms.

3 Corrosion detection approach

3.1 Motivation

The Yolov3-tiny (Redmon and Farhadi 2018) is an object detector which has been proved to be fast and accurate on embedded platforms. Fig. 1 shows the network structure of the Yolov3-tiny. There are seven convolution layers and six max pooling (MaxPool) layers for extracting image features. Two-scale detection is utilised to detect different scale targets. The detection process of the Yolov3-tiny is described as follows:

Step 1   Load the input image and resize the image into size 416*416

Step 2   Extract features with convolutional and MaxPool layers

Step 3   Produce feature maps of size 13*13 on a small scale

Step 4   Upsample small scale feature maps to size 26*26 and connect them to the same size feature maps generated by the feature extraction network

Step 5   Produce feature maps of size 26*26 on a large scale.

Step 6   Divide the input image into 13*13 and 26*26 grids for two-scale object detection. Based on the predefined anchors, the grid will be responsible for predicting the object when the centre of the object lies in the grid.

Step 7   Output the two-scale prediction results

Step 8   Fuse different scale prediction results and acquire accurate bounding boxes

Since the simple and shallow network is designed as the backbone, detection accuracy of the Yolov3-tiny is not high enough (Fang et al. 2019). Moreover, the Yolov3-tiny deploys many convolution layers with 512 and 1024 convolution filters, which leads to a large number of parameters of the network and requires enormous storage space for embedded platforms.

To address these problems in the Yolov3-tiny for corrosion detection, this paper proposes a novel metallic corrosion detector. The overall schematic architecture is presented in Fig. 2. A brand-new lightweight backbone network with the DSConv (Howard et al. 2017) to reduce parameters of the model is designed. Considering the simplified backbone network cannot extract robust corrosion features, as a complement, the SPP is introduced and improved to fuse local features. Meanwhile, the attention mechanism, three-scale object detection and focal loss are adopted to further improve the feature extraction capabilities and prediction accuracy. Details of the AMCD will be explained in the following parts.

3.2 Structure of the AMCD

As shown in Fig. 2, the backbone is responsible for extracting features from images, and the detection part will output the position and category of the corrosion. As the AMCD focuses on achieving accurate corrosion detection on embedded platforms, a shallow network has been designed. What

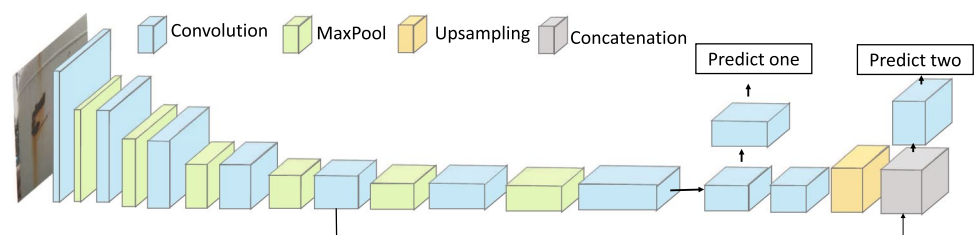**Fig. 1** Framework of the Yolov3-tiny

Fig. 2 Structure of the AMCD

Backbone					Detection
Type	Filters	Size/Stride	Output		

Backbone:

Type	Filters	Size/Stride	Output
Conv	16	3x3/2	160
DSConv	32	3x3/2	80
DSConv	64	3x3/2	40
DSConv	64	3x3/1	40
Concat			40
CBAM			40
DSConv	128	3x3/2	20
DSConv	128	3x3/1	20
Concat			20
CBAM			20
SPP			10
CBAM			10
DSConv	256	3x3/1	10
DSConv	256	3x3/1	10
Concat			10
CBAM			10

Detection:
Concat → DSConv 128 → Conv 2D (Prediction three)
Upsample
DSConv 128
Concat → DSConv 256 → Conv 2D (Prediction two)
Upsample
DSConv 256 → Conv 2D (Prediction one)
DSConv 512

(a) DSConv

FM_in   k x k   Depthwise Convolution   Pointwise Convolution   1 x 1

(b) Standard convolution
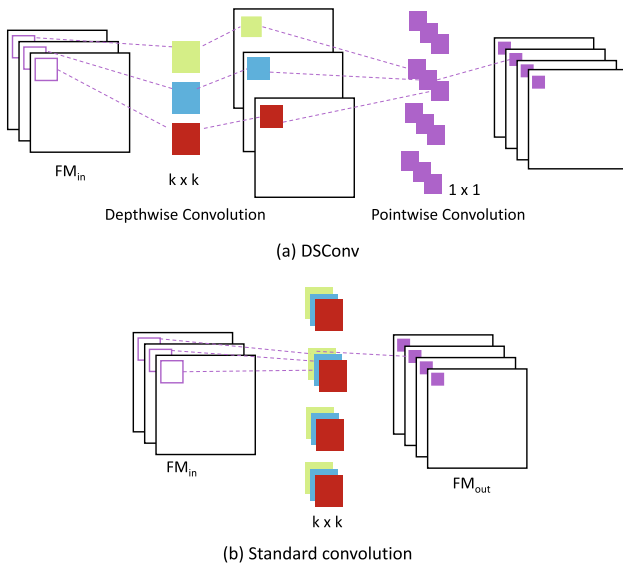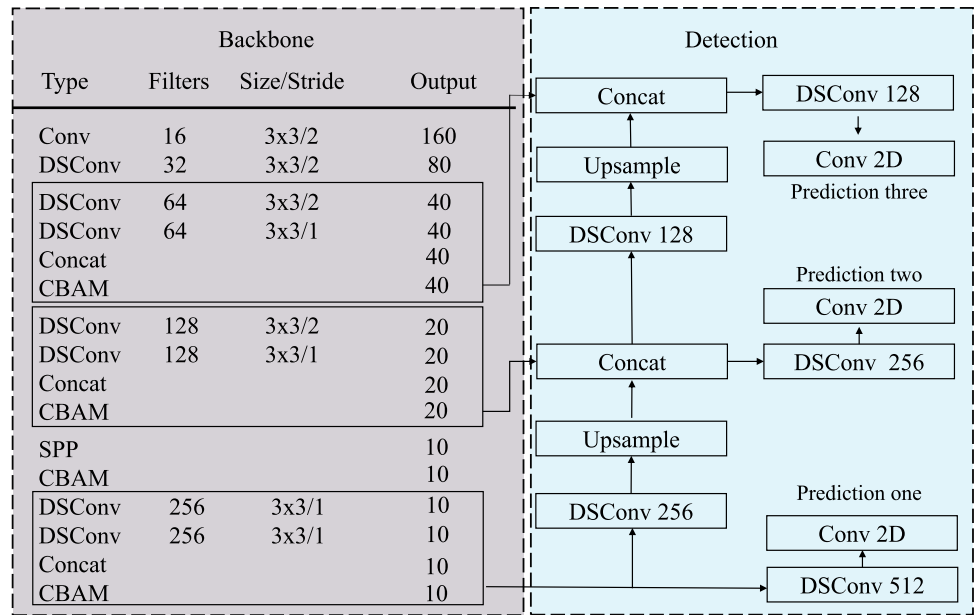
FM_in   k x k   FM_out

Fig. 3 Comparison of the DSConv with standard convolution

is more, the DSConv is adopted to reduce parameters. To enhance the feature extraction capability of the shallow network, the CBAM (Woo et al. 2018), three-scale prediction, improved SPP and focal loss (Lin et al. 2017) are also utilised. Finally, the designed backbone contains 1 traditional convolution layer, 7 DSConv layers, 1 SPP layer and 4 CBAM modules.

DSConv The DSConv can reduce model parameters significantly. Unlike the traditional convolution processes images from height, width and channel dimensions simultaneously, the DSConv divides the convolution process into the depthwise convolution and pointwise

convolution. Fig. 3 demonstrates the comparison of the DSConv with standard convolution.

The first step of the DSConv is a depthwise convolution. In this step, the number of filters is the same as that of input channels, which ensures that only one feature map is generated through one input channel. The equation of the depthwise convolution is shown as follows:

$$DWConv(W, X)_{(i,j)} = \sum_{m,n}^{M,N} W(m, n) \cdot X_{(i+m,j+n)} \quad (1)$$

where W indicates the weight matrix of convolutional filters. X denotes input feature maps. (i, j) represents the coordinate of point within output feature maps. M and N are the height and width of input feature maps. Meanwhile, m and n represent the height and width of the convolutional filter, respectively.

In the second step, a set of 1×1 convolutional layers is applied to fuse feature maps generated by the depthwise convolution, and it is called the pointwise convolution. The pointwise convolution focuses on the combination of spatial features, which only changes the number of channels while keeping the width and height of feature maps. The formula of the pointwise convolution is written as:

$$PWConv(W, x)_{(i,j)} = \sum_{c}^{C} W(c) \cdot x_{(i,j,c)} \quad (2)$$

where C is the total number of channels of input feature maps. c represents the channels of convolution filters.

Overall, the whole process can be represented by:

$$DSConv(W_d, W_p, X)_{(i,j)} =$$
$$PWConv(W_p, DWConv(W_d, X)_{(i,j)})_{(i,j)} \qquad (3)$$

At the same time, the formula of the standard convolution can be represented by:

$$Conv(W, X)_{(i,j)} = \sum_{m,n,c}^{M,N,C} W(m, n, c) \cdot X_{(i+m,j+n,c)} \qquad (4)$$

Compared with the traditional convolution, parameters of the DSConv are reduced significantly. There is an assumption that the number of output channels is o. According to Eq. (4), total parameters of the standard convolution are $m \times n \times c \times o$. While the DSConv is utilised to generate the same output feature maps, based on Eq. (3), total parameters are $m \times n \times c + c \times o$. Comparison of parameters between the DSConv and standard convolution is presented:

$$\frac{m \times n \times c + c \times o}{m \times n \times c \times o} = \frac{1}{o} + \frac{1}{m \times n} \qquad (5)$$

For example, input feature maps contain 3 channels. The convolutional kernel size is 3×3, and there are 4 sets of convolutional filters to output 4 feature maps. The standard convolution processes images from height, width and channel dimensions simultaneously, and all parameters are 108. Meanwhile, same input feature maps are processed by the DSConv to output same size and channels of feature maps. At first, every single channel of the input feature map is processed by a 3×3 convolutional filter. Then, 4 sets of $1 \times 1 \times 3$ convolutional filters are utilised to process the generated feature map and output 4 feature maps. The total parameters of the DSConv are 39, which is almost 1/3 of the traditional convolution.

SPP The Yolov3-tiny only focuses on the fusion of global features extracted by different scale convolutional layers (Huang et al. 2020). To take the concatenating of local region features on the same convolutional layer, a new SPP module is designed. The combination of global and local features is utilised to improve the performance of corrosion detection.

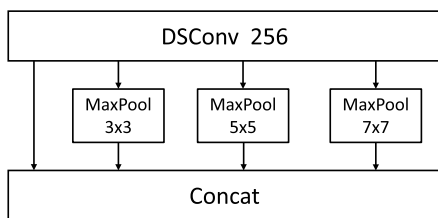The architecture of the SPP presented by this paper is shown in Fig. 4. Different from the traditional SPP proposed

by He et al. (2015b), the improved SPP does not resize feature maps into feature vectors. Instead, the improved SPP outputs feature maps. Based on the size of input feature maps, MaxPool layers with kernel sizes of 3×3, 5×5 and 7×7 are utilised to pool feature maps. The stride of each pooling layer is 1, and padding is adopted to make sure the size of generated feature maps is the same as that of input feature maps. After the concatenation, there are 1024 feature maps generated by the improved SPP which extracts and fuses local region features.

CBAM Inspired by human visual attention mechanism, CNNs can employ attention mechanism to select optimal information from the training dataset. The attention module selects the most representative area in the image and allows the network to focus on there. Thus, more critical features can be extracted, and the detection accuracy will be improved. The attention mechanism has proven its effectiveness in many tasks, such as river detection (Gao et al. 2017b), outdoor illumination estimation (Jin et al. 2020) and synthetic aperture radar (SAR) image recognition (Gao et al. 2019).

The CBAM outputs refined feature maps by channel and spatial attention sequentially. The overview diagram of the CBAM is shown in Fig.5. In general, the channel attention module focuses on figuring out optimal feature maps between different channels of feature maps. The spatial attention module aims to output a spatial attention map
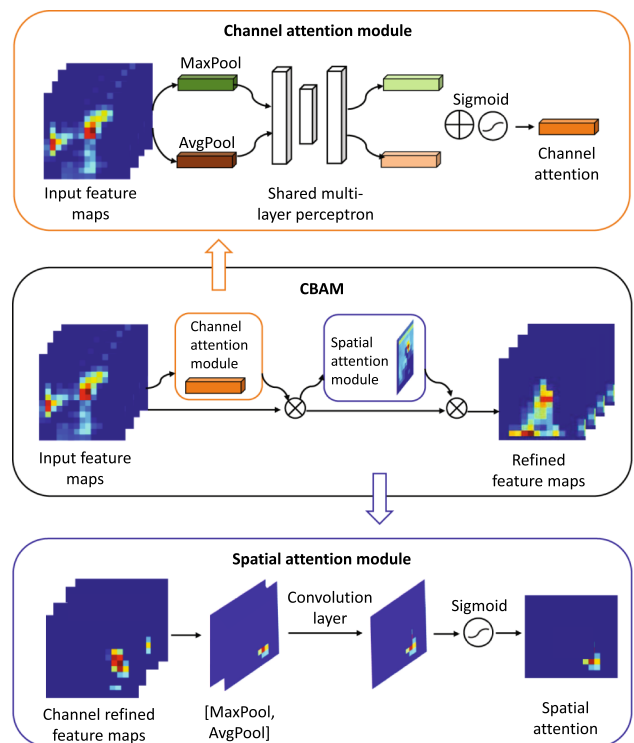

Fig. 4 Structure of the improved SPP

Fig. 5 Diagram of the CBAM

based on local information. The MaxPool and global average pooling (AvgPool) operations are utilised to construct feature map statistics. The MaxPool could return the significant features of the target. At the same time, the AvgPool provides global statistics of feature maps. With the usage of MaxPool and AvgPool operations, the representation of features extracted by CNNs is improved. Channel attention focuses on global information, whereas spatial attention is employed locally. Therefore, the CBAM can extract comprehensive salient features to improve the performance of corrosion detection.

3.3 Loss function

The Yolov3-tiny uses anchors to generate candidate object location from the whole image. The number of potential bounding boxes containing objects is much less than those only containing background. What is more, negative samples contribute no useful learning signal and cause biased learning. Finally, it will lead to a degenerate detector, which cannot detect the corrosion correctly. To overcome this limitation, the focal loss is introduced into the AMCD, which gives a high loss value to an object. This makes the detector concentrate on object areas and become sensitive to the target. The formula of the focal loss is:

$$F = \alpha \times (1 - p)^{\lambda} \tag{6}$$

The α is the hyperparameter which down-weights the loss contributed by backgrounds. p indicates the confidence of whether the candidate bounding box contains the object. λ represents the exponential scaling factor which down-weights the loss generated by easy examples and makes the CNNs focus on difficult examples. In the AMCD, the α and λ are 0.5 and 2, respectively.

According to Eq.(6), the loss function of the AMCD can be defined as:

$$
\begin{aligned}
Loss = & \\
& \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} [(x_i - \widehat{x}_i)^2 + (y_i - \widehat{y}_i)^2] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^{B} 1_{ij}^{obj} (2 - w_i \times h_i)[(w_i - \widehat{w}_i)^2 + (h_i - \widehat{h}_i)^2] \\
& + F \sum_{i=0}^{S^2} \sum_{j=0}^{A} 1_{ij}^{obj} [\widehat{C}_i log(C_i) + (1 - \widehat{C}_i) log(1 - C_i)] \\
& + F \sum_{i=0}^{S^2} \sum_{j=0}^{A} 1_{ij}^{noobj} [\widehat{C}_i log(C_i) + (1 - \widehat{C}_i) log(1 - C_i)] \\
& + \sum_{i=0}^{S^2} 1_{ij}^{obj} \sum_{c \in classes} [\widehat{p}_i(c) log(p_i(c)) + (1 - \widehat{p}_i(c)) log(1 - p_i(c))]
\end{aligned}
\tag{7}
$$

where S^2 denotes the number of grid cells. B is the number of bounding boxes predicted by each cell. 1_{ij}^{obj} indicates that the jth bounding box predicted in cell i contains an object. At the same time, 1_{ij}^{noobj} refers to the predicted bounding box only contains background. x_i and y_i are the centre coordinate of the bounding box. w and h represent the dimension of the bounding box. The variables with $\widehat{}$ indicate they are predicted values. Otherwise, they are groundtruth. C_i denotes the confidence of whether the bounding box contains an object or just pure background. The prediction of classes is represented by $pi(c)$.

4 Experiments and analysation

4.1 Dataset

To construct a dataset to train and verify the AMCD, 5625 images are captured by a DJI Phantom 4. Images are taken from different facilities such as pressure vessels and oil wells at a distance from 1m-10m under different angles and illumination conditions. These images contain the bar corrosion, nubby corrosion, fastener corrosion and exfoliation. If the aspect ratio of a damage area is less than 1:2, this region will be treated as the nubby corrosion. Otherwise, the damage region will be considered as the bar corrosion. Bolt and nut corrosions are treated as the fastener corrosion, and the exfoliation corrosion includes cracked coatings. To annotate captured images with different corrosion types, labelImg (https://tzutalin.github.io/labelImg/) is utilised to put bounding boxes on images by human experts. Each bounding box contains the upper-left corner position and width and height of the box. Thus the format of the bounding box is (x, y, w, h). There is a total of 27039 corrosion areas labelled in 5625 images. Several annotated images are shown in Fig. 6. Bounding boxes with different colours represent different kinds of corrosions.

To generate training and test sets, labelled images are randomly divided by contained corrosions. The training and validation dataset contains 4500 images. Other 1125 images are utilised to test the proposed detector.

4.2 Experimental environment

All training processes are conducted by applying Tensorflow 1.15 and CUDA 10.0 on a computer with an Intel®CoreTM i7-8750 @2.20 GHz CPU, 12 GB installed memory (RAM) and 6 GB GDDR5 memory NVIDIA GTX1060 graphics processing unit (GPU). To evaluate the performance of the AMCD on MAVs on-board platform, testing processes are made on the Nvidia Jetson TX2. It is equipped with a hexacore CPU and an NVIDIA Pascal™ family GPU with 256

CUDA cores. It loads with 8GB of memory and 59.7GB/s of memory bandwidth.

Transfer learning has the ability to transfer knowledge from a related task that has already been learned to a new domain. A lot of works have proved that transfer learning is an optimisation technique which saves training time and gets better test performance. Instead of using randomly initialised weights of CNNs, layers of the proposed model are initialised by the weight trained on PASCAL VOC2007 (Everingham et al. 2007) and PASCAL VOC2012 (Everingham and Winn 2011) datasets. To further improve the performance of the ACMD, predefined anchors are clustered by K-means (Kanungo et al. 2002) as [14,15, 18,21, 31,17, 25,26, 20,38, 36,35, 30,78, 63,48, 93,118].

The explained network is trained using the SGD algorithm, and the tuning of learning rates influences the training performance significantly. To make the training process stable and efficient, warmup (He et al. 2016) and cosine learning rate decay (Loshchilov and Hutter 2016) are utilised. Fig. 7 depicts the variation of the learning rate during the training stage. The x axis represents iterations, and the learning rate is updated every iteration. The momentum parameter is 0.9995. The batch size is assigned by 2, and the model is trained for 450000 iterations. The loss curve during the training process can be seen in Fig. 8, which indicates the loss function is optimised and convergent to a stable value.

4.3 Evaluation metrics

Precision and recall concepts (Olson and Delen 2008) are utilised widely to evaluate the performance of object detection approaches. Precision denotes the number of
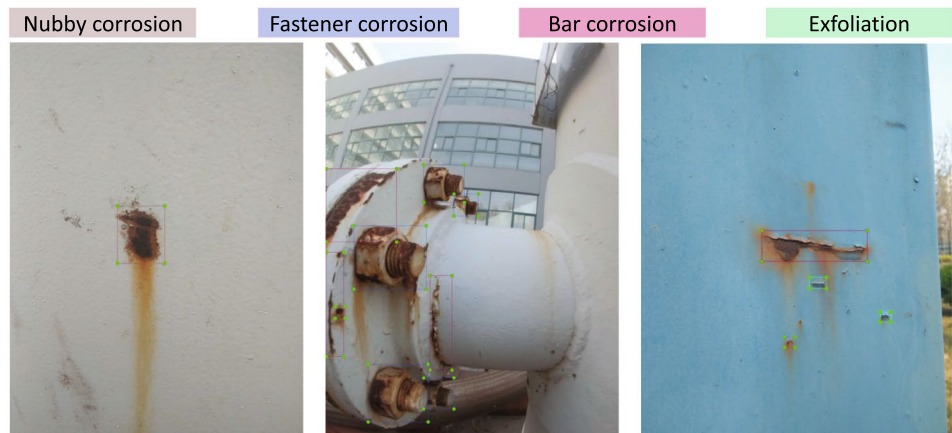
Fig. 6 Labelled images

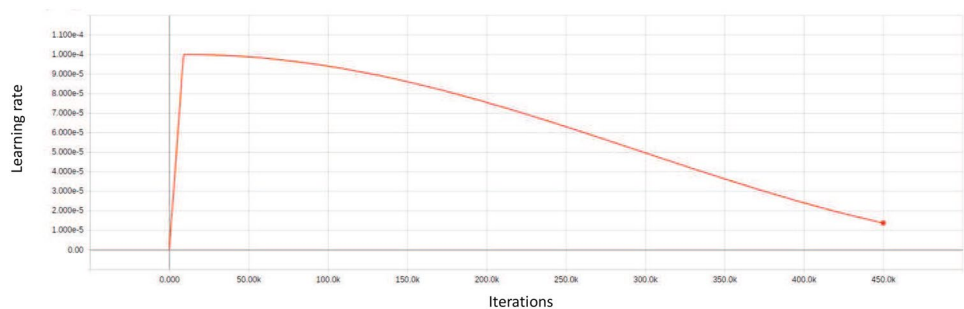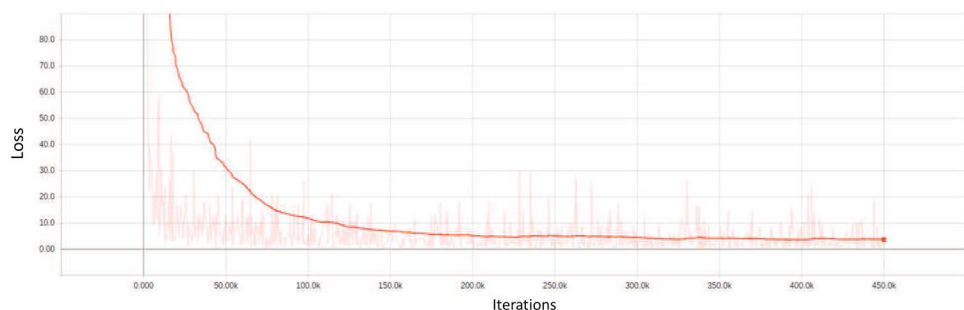**Fig. 7** Learning rate curve during the training procedure

Fig. 8 Loss decline curve of the AMCD

True Positive (TP) results divided by all positive detection results. Recall is defined as the percentage of the TP in all correct detection results. The area under the precision and recall curve is called average precision (AP). The AP indicates the ability of the detector to locate objects and classify them into a single class. In general, the higher AP for a category of objects, the better performance of the detector to identify them. Mean average precision (mAP) represents the performance of the detector across all classes and can be defined by the average value of APs for all classes.

In reality, predicted results cannot match groundtruth perfectly. Thus, the Intersection-over-Union (IoU) metric is adopted to represent the overlap of the predicted bounding box with groundtruth box. This allows predicted results can partially overlap with groundtruth. While the overlap area between suspicious corrosion and groundtruth exceeds the IoU threshold, the prediction result is classified as positive. Otherwise, the detection result is categorised as negative. In this study, the IoU value is 0.5.

4.4 Performance of the AMCD

The trained model is used to identify different kinds of corrosions, and some recognition results are shown in Fig. 9. As images are taken under different angles and illumination conditions, backgrounds are cluttered. We can see that four kinds of corrosions can be detected correctly. What is more, when an image contains multiple types of corrosions, all of corrosions can be identified. As shown in Fig. 9b, the fastener corrosion and bar corrosion are detected correctly, even though some shadows exist in the image. In Fig. 9f, small holes in the structure are very similar to the nubby corrosion, the AMCD still can locate corrosion areas precisely.

Precision-recall (PR) curves and APs for four kinds of corrosions are demonstrated in Fig. 10 and Table 1, respectively. Based on unique features of the fastener corrosion, the detection results show a high accuracy towards this kind of corrosion. As its shape distinguishes bar corrosion and nubby corrosion, the features extracted from them are similar. The number of the nubby corrosion in the training dataset is far more than that of the bar corrosion. Thus, the bar corrosion can be misunderstood as the nubby corrosion easily, leading to a relatively low detection accuracy of the

Fig. 9 Some detection results

(a)   (b)   (c)
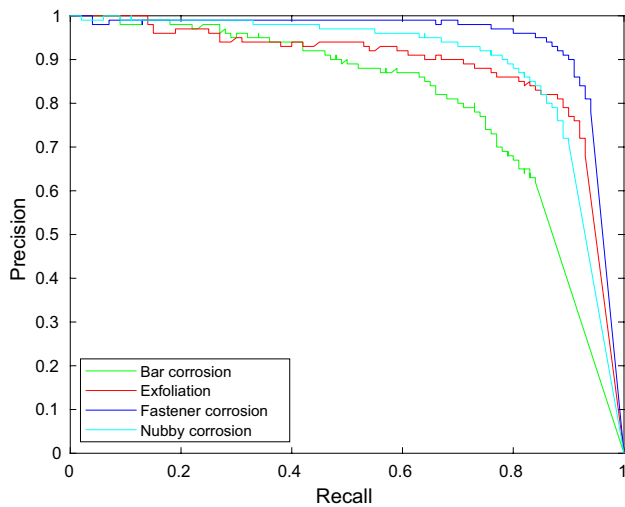
(d)   (e)   (f)

Fig. 10 PR curves of four kinds of corrosions

Table 1 Detection results of four kinds of corrosions

	Number of corrosions	Detection accuracy (%)	AP (%)
Nubby corrosion	3559	89.83	85.81
Bar corrosion	501	84.23	75.78
Exfoliation	290	93.45	86.36
Fastener corrosion	985	93.81	91.88

bar corrosion. Besides its unique features, the exfoliation contains part of features of the nubby corrosion or bar corrosion. Due to the least samples of the exfoliation gathered, its detection accuracy is between that of the bar corrosion and nubby corrosion.

4.5 Comparison with state-of-the-art detection methods

In this section, the proposed network is compared with some state-of-the-art detectors. As the AMCD towards to detect corrosion with limited computing resources, the Yolov2-tiny (Redmon and Farhadi 2017), Yolov3-tiny, Yolov4-tiny (Bochkovskiy et al. 2020), single shot multibox detector (SSD) (Liu et al. 2016) and RetinaNet (Lin et al. 2017) are selected for the comparison. To make a fair comparison, the input dimension of compared detectors is resized into a similar scale, and all compared detectors are trained with default parameters provided by the authors. As shown in Table 2, the proposed detector achieves 84.96% mAP, which is the best among these algorithms. The superiority of the AMCD is validated.

Table 2 Comparison of corrosion detection performance

Detector	Backbone	Input dimension	mAP (%)
Yolov2-tiny	Yolov2-tiny	320	71.87
Yolov3-tiny	Yolov3-tiny	320	79.02
Yolov4-tiny	Yolov4-tiny	320	82.1
SSD	VGG16	300	81.2
RetinaNet	Resnet50	320	83.5
AMCD (ours)	AMCD	320	84.96

Figure 11 shows that the proposed method can achieve optimal corrosion detection results compared with other state-of-the-art algorithms. Other detectors are limited by the tightly layout and sizes of corrosions. The Yolov2-tiny is struggling to generate accurate bounding boxes to corrosion areas. The Yolov3-tiny, Yolov4-tiny and RetinaNet cannot detect small-size corrosions. The SSD and AMCD identify corrosions correctly in this image. When taking the model size and detection speed into consideration, the AMCD still outperforms the SSD. More details are shown in Fig. 12. Due to the shallow network architecture and DSConv being utilised in the AMCD, the detection speed reaches 20.18 FPS on average. It is much faster than the SSD, which uses the VGG16 as the backbone. With the adoption of the DSConv, the model size reduces significantly, and it is only 6.1 MB. This suggests that the AMCD could perform corrosion detection efficiently, which is essential for MAVs onboard visual inspection applications.

5 Conclusions

In this study, a deep learning-based corrosion detection technique has been developed for automated visual inspection of steel structures in industrial areas with MAVs. This study focuses on processing images captured by MAVs to identify the presence of corrosions on structures with limited computing resources. In the step of detection, we use the DSConv to build the backbone, which can significantly reduce parameters and improve the detection speed. In order to improve the detection accuracy of the AMCD, the attention mechanism, three-scale object detection and focal loss are adopted, which are helpful for accurate corrosion detection. What is more, the improved SPP is introduced for further improving the detection accuracy. The proposed approach achieves excellent performance in detecting and recognising different categories of corrosions. Experimental results prove that the proposed detector obtains satisfactory corrosion detection results, which is able to achieve 84.96% mAP for corrosion identification in the complex environment and get the real-time
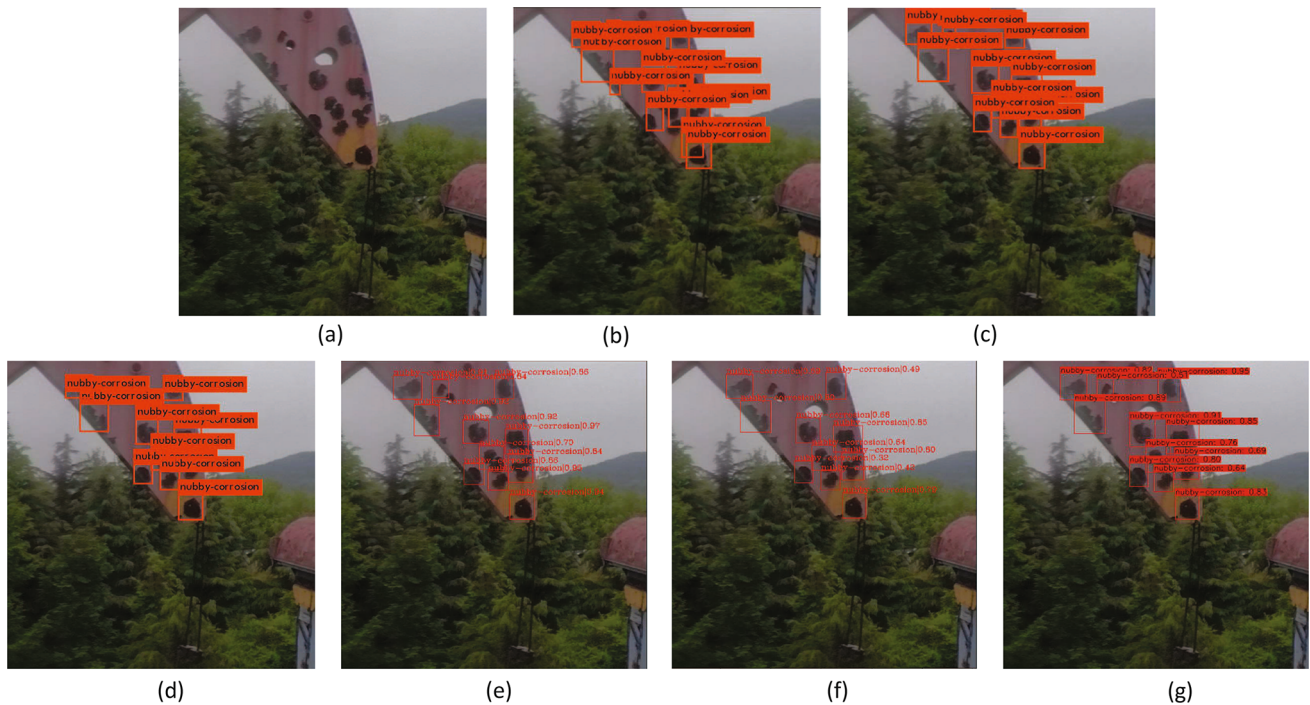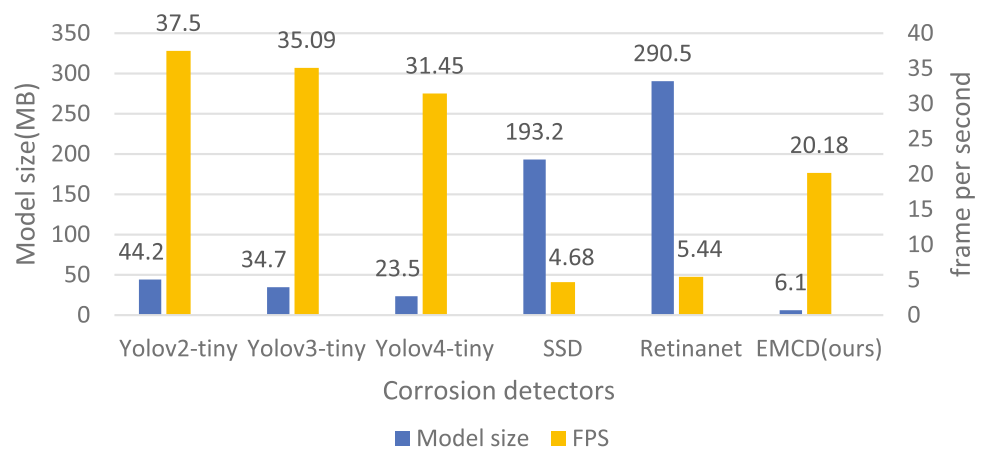
Fig. 11 Original image (**a**) and detection results produced by the Yolov2-tiny (**b**), Yolov3-tiny (**c**), Yolov4-tiny (**d**), SSD (**e**), RetinaNet (**f**) and AMCD (**g**)

Fig. 12 Comparison of the Yolov2-tiny, Yolov3-tiny, Yolov4-tiny, SSD, RetinaNet and AMCD in terms of model sizes and FPS

performance (20.18 FPS) with off-the-shelf MAVs commercial on-board processing platform. What is more, the model size is only 6.1MB.

Author Contributions Idea conception, LY; project supervision, EY; formal analysis, LY and EY; dataset construction, CL and PR; original draft writing, LY; review and editing, EY, CL and PR.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Agnisarman S, Lopes S, Madathil KC, Piratla K, Gramopadhye A (2019) A survey of automation-enabled human-in-the-loop systems for infrastructure visual inspection. Autom Constr 97:52–76

Atha DJ, Jahanshahi MR (2018) Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. Struct Health Monitor 17(5):1110–1128

Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. arXiv:200410934

Bonnín-Pascual F, Ortiz A (2010) Detection of cracks and corrosion for automated vessels visual inspection. In: CCIA, pp 111–120

Bonnin-Pascual F, Ortiz A (2014) Corrosion detection for automated visual inspection. In: Developments in corrosion protection, IntechOpen

Cha YJ, Choi W, Suh G, Mahmoudkhani S, Büyüköztürk O (2018) Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Comput Aided Civil Infrastruct Eng 33(9):731–747

Chen Q, Wen X, Lu S, Sun D (2019) Corrosion detection for large steel structure base on uav integrated with image processing system. In: IOP Conference Series: Materials Science and Engineering, IOP Publishing, vol 608, p 012020

Chen W, Qiao Y, Li Y (2020) Inception-ssd: an improved single shot detector for vehicle detection. J Ambient Intell Humaniz Comput. https://doi.org/10.1007/s12652-020-02085-w

Chu M, Thuerey N (2017) Data-driven synthesis of smoke flows with cnn-based feature descriptors. ACM TOG 36(4):1–14

Du J, Yan L, Wang H, Huang Q (2018) Research on grounding grid corrosion classification method based on convolutional neural network. In: MATEC web of conferences, EDP Sciences, vol 160, p 01008

Everingham M, Winn J (2011) The pascal visual object classes challenge 2012 (voc2012) development kit. Pattern Anal Stat Modell Comput Learn. https://doi.org/10.1007/s11263-009-0275-4

Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A (2007) The pascal visual object classes challenge 2007 (voc2007) results. Int J Comput Vis 88:303–338

Fang W, Wang L, Ren P (2019) Tinier-yolo: a real-time object detection method for constrained environments. IEEE Access 8:1935–1944

Gao F, Huang T, Wang J, Sun J, Hussain A, Yang E (2017a) Dual-branch deep convolution neural network for polarimetric sar image classification. Appl Sci 7(5):447

Gao F, Ma F, Wang J, Sun J, Yang E, Zhou H (2017b) Visual saliency modeling for river detection in high-resolution sar imagery. IEEE Access 6:1000–1014

Gao F, Shi W, Wang J, Hussain A, Zhou H (2019) A semi-supervised synthetic aperture radar (sar) image recognition algorithm based on an attention mechanism and bias-variance decomposition. IEEE Access 7:108617–108632

Gomes WJ, Beck AT, Haukaas T (2013) Optimal inspection planning for onshore pipelines subject to external corrosion. Reliab Eng Syst Safety 118:18–27

Guindel C, Martín D, Armingol JM (2017) Modeling traffic scenes for intelligent vehicles using cnn-based detection and orientation estimation. In: Iberian Robotics conference, Springer, Berlin, pp 487–498

He K, Zhang X, Ren S, Sun J (2015a) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp 1026–1034

He K, Zhang X, Ren S, Sun J (2015b) Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE Trans Pattern Anal Mach Intell 37(9):1904–1916

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 770–778

Hoang ND (2020) Image processing-based pitting corrosion detection using metaheuristic optimized multilevel image thresholding and machine-learning approaches. Math Probl Eng. https://doi.org/10.1155/2020/6765274

Hoang ND, Tran VD (2019) Image processing-based detection of pipe corrosion using texture analysis and metaheuristic-optimized machine learning approach. Comput Intell Neurosci. https://doi.org/10.1155/2019/8097213

Hoskere V, Narazaki Y, Hoang T, Spencer Jr B (2018) Vision-based structural inspection using multiscale deep convolutional neural networks. arXiv:180501055

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:170404861

Huang Z, Wang J, Fu X, Yu T, Guo Y, Wang R (2020) Dc-spp-yolo: dense connection and spatial pyramid pooling based yolo for object detection. Inform Sci. arXiv:1903.08589

Jin X, Deng P, Li X, Zhang K, Li X, Zhou Q, Xie S, Fang X (2020) Sun-sky model estimation from outdoor images. J Ambient Intell Human Comput. https://doi.org/10.1007/s12652-020-02367-3

Kanungo T, Mount DM, Netanyahu NS, Piatko CD, Silverman R, Wu AY (2002) An efficient k-means clustering algorithm: analysis and implementation. IEEE Trans Pattern Anal Mach Intell 24(7):881–892

Khaire UM, Dhanalakshmi R (2019) Stability of feature selection algorithm: a review. J King Saud Univ Comput Inform Sci. :https://doi.org/10.1016/j.jksuci.2019.06.012

Kumar V, Recupero DR, Riboni D, Helaoui R (2021) Ensembling classical machine learning and deep learning approaches for morbidity identification from clinical notes. IEEE Access 9:7107–7126

Li J, Su Z, Geng J, Yin Y (2018) Real-time detection of steel strip surface defects based on improved yolo detection network. IFAC 51(21):76–81

Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp 2980–2988

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: single shot multibox detector. In: European Conference on Computer Vision, Springer, Berlin, pp 21–37

Loshchilov I, Hutter F (2016) Sgdr: Stochastic gradient descent with warm restarts. arXiv:160803983

Moosavi AN (2017) Corrosion in onshore production and transmission sectors–current knowledge and challenges. In: Trends in oil and gas corrosion research and technologies, Elsevier, Amsterdam, pp 95–109

Olson DL, Delen D (2008) Performance evaluation for predictive modeling. In: Advanced data mining techniques, Springer, Berlin, pp 137–147

Pereira MC, Silva JW, Acciari HA, Codaro EN, Hein LR (2012) Morphology characterization and kinetics evaluation of pitting corrosion of commercially pure aluminium by digital image analysis. Mater Sci Appl 3:7

Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7263–7271

Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv:180402767

Sarkar D, Gunturi SK (2020) Wind turbine blade structural state evaluation by hybrid object detector relying on deep learning models. J Ambient Intell Humaniz Comput 12:8535–8548

Tscheliesnig P, Jagenbrein A, Lackner G (2016) Detecting corrosion during inspection and maintenance of industrial structures using acoustic emmision. In: 32nd European Conference on Acoustic Emission Testing, pp 7–9

Woo S, Park J, Lee JY, So Kweon I (2018) Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 3–19

Xu H, Su X, Wang Y, Cai H, Cui K, Chen X (2019) Automatic bridge crack detection using a convolutional neural network. Appl Sci 9(14):2867

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.