Small Foreign Metal Objects Detection in X-Ray Images of Clothing Products Using Faster R-CNN and Feature Pyramid Network

Rong Gao, Zhaoyun Sun, Ju Huyan, Wei Li, Liyang Xiao, Bobin Yao, *Member, IEEE*, and Huifeng Wang

***Abstract*— Immediate and accurate detection of foreign metal objects (FMOs) in clothing products is important for guaranteeing human safety. This article proposes an online detection approach based on deep learning, which is suitable for detecting small FMOs from X-ray images of clothing packages. A conveyor belt X-ray scanning system is developed for image collection. The X-ray images are preprocessed by using the morphological erosion operation to improve the accuracy of FMOs detection. These images are then down-sampled to reduce the computation cost. Feature pyramid network (FPN) is adopted for aggregating feature maps with different resolutions, which proved to be effective for small FMOs detection. The stochastic gradient descent (SGD) is used to optimize a multitask loss. The trained model was tested offline on 200 X-ray images, which achieved precision = 0.999, recall = 0.988, F1-score = 0.993, and AP = 0.946. Compared to original Faster region-based convolutional neural networks (R-CNN), the proposed method significantly improved the performance for small FMOs detection in terms of precision and recall rate.**

***Index Terms*— Clothing product, convolutional neural network (CNN), foreign metal object (FMO) detection, X-ray.**

I. Introduction

THE occurrence of foreign metal objects (FMOs) such as needles, staples, or wires in packaged clothing products can cause serious dissatisfactory to both the manufacturer and consumer. Therefore, detection of FMOs is an essential part for the quality control of clothing products to prevent consumers from suffering from physical or mental harm.

Manuscript received February 19, 2021; revised April 19, 2021; accepted April 27, 2021. Date of publication May 14, 2021; date of current version May 25, 2021. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1600202, in part by the National Natural Science Foundation of China under Grant 51978071 and Grant 61601058, in part by the Fundamental Research Funds for the Central Universities under Grant 300102249301 and Grant 300102249306, and in part by the Project of He'nan Transportation Science and Technology Plan under Grant 2021G8. The Associate Editor coordinating the review process was Lihui Peng. *(Corresponding authors: Zhaoyun Sun; Ju Huyan.)*

Rong Gao, Zhaoyun Sun, Wei Li, and Liyang Xiao are with the School of Information Engineering, Chang'an University, Xi'an 710064, China (e-mail: gr@chd.edu.cn; zhaoyunsun@126.com; grandy@chd.edu.cn; liyang_xiao@163.com).

Ju Huyan is with the School of Transportation, Southeast University, Nanjing 211189, China (e-mail: ju_huyan@outlook.com).

Bobin Yao and Huifeng Wang are with the School of Electronics and Control Engineering, Chang'an University, Xi'an 710064, China (e-mail: b.b.yao@chd.edu.cn; hfwang@chd.edu.cn).

Digital Object Identifier 10.1109/TIM.2021.3077666

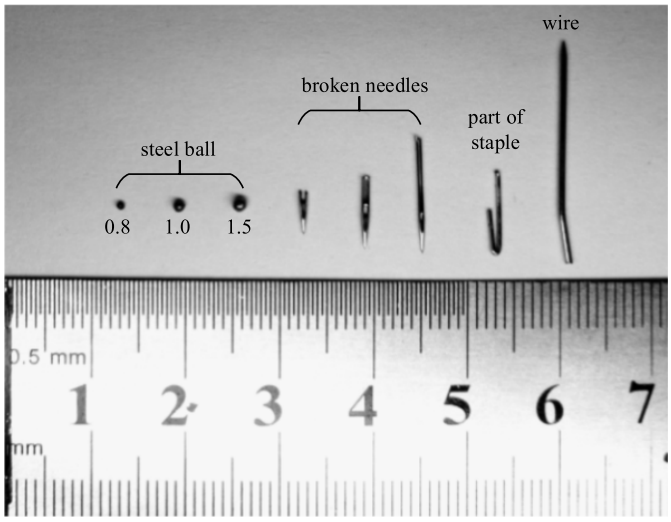

Fig. 1. FMOs samples for the experiment.

As a contactless vision-based sensor [1], an X-ray scanner works by safely passing low-energy X-rays into the package as it passes through the detector. Each clothing item can absorb a specific frequency range of X-ray energy, including the FMOs. The higher an object's density, the harder it is for X-rays to pass through. The FMOs absorb more X-ray energy than regular clothing areas surrounding them, which usually resulted in appearing a black area in X-ray images. During the production and packaging procedures, FMOs can be mixed into the packaging box because of several reasons. As an example, Fig. 1 shows four types of most commonly seen FMOs collected in clothing factories. The shapes and dimensions of these FMOs are quite different from one another. However, one common feature among all these FMOs is that they are all small objects that are hard to be detected. It is time-consuming and labor-intensive to identify small FMOs in the X-ray images by human vision. Therefore, production lines must deploy intelligent detection systems.

This work treated various types of small FMOs as a single class, and focused on detecting and localizing FMOs from the X-ray images. As a vision-based measurement (VBM) [2] application, objects detection from X-ray images has application prospects in weld defects detection [3], nondestructive testing (NDT) [4], [5], luggage screening [6], and

TABLE I
TOP SUBMISSIONS FOR MS COCO 2020 OBJECT DETECTION

RANK	MODEL	BOX AP	APS	APL
1	CSP-p7+Mish	55.8	38.8	68.2
2	YOLOv4-P7	55.4	38.1	67.4
3	EfficientDet-D7x	55.1	37.2	68.0
4	CSP-p6+Mish	54.9	37.4	66.7

TABLE II
DEFINITIONS OF THE SMALL, MEDIUM, AND LARGE OBJECTS IN MS COCO

CLASS	Min box	Max box	Proportion in image
Small	0×0	32×32	0%~5%
Medium	32×32	96×96	5%~15%
Large	96×96	∞×∞	15%~100%

computer-aided diagnostic (CAD) [7]–[18]. The commonly used object detection method can be grouped into two categories, which are conventional method and artificial intelligence (AI)-based methods. Conventional object detection methods mainly focused on extracting different types of hand-crafted features rely on experts. Meanwhile, complex image preprocessing operations are required to improve object detection results. In recent years, deep learning, especially convolutional neural network (CNN), has achieved remarkable success in solving diverse complex computer vision-based problems, the field of object detection is no exception.

A challenging problem which arises in this domain is the small object detection. Small object detection is essential for many downstream applications. Safety of autonomous vehicles requires the long-distance detection of small targets, such as traffic lights or pedestrians [19], from high-resolution images. In medical images, early detection of masses or tumors [20] is critical for accurate diagnosis, among which targets are likely to be only a few pixels in size. Other conditions, such as satellite images, the target objects of cars [21], ships [22], and buildings [23] with an average pixel size representing a resolution of 0.5–5 m may cover only a few pixels. However, to detect small objects, the performances of the current state-of-the-art object detection method are far from satisfactory, especially under the real-time requirements. Top four object detection models on the Microsoft Common Objects in Context (MS COCO) dataset are shown in Table I. It can be seen that the average precision (AP) of small object detection is only half of that of large objects.

Table II lists the definitions of objects scale in MS COCO. Objects with a maximum bounding box area of less than 32×32 are considered as small objects. The images with the maximum resolution in the COCO dataset are 640×640. In other words, if an object accounts for less than 5% of the total image pixels, it will be considered as a small object according to MS COCO definition. In Fig. 1, the smallest FMOs are steel balls with a diameter of 0.8 mm, while the width of a staple is only 0.4 mm. Meanwhile, the clothes packaging box is usually 300–600 mm in width and 400–900 mm in length. To detect those types of FMOs from packaging boxes in the production lines, the collected X-ray images should contain the entire packaging box, while the necessary resolution to distinguish the smallest FMOs should be maintained as well. An image with at least 2250×1500 pixels is needed for scanning a clothing package of 900 mm (length) $\times$ 600 mm (width), which can ensure a resolution of 0.4 mm. In this situation, a steel ball with a diameter of 0.8 mm in the image only represented by 9–16 pixels, meaning that only around 0.00027%–0.00047% of the entire X-ray image are the target pixels. It is extremely small under the MS COCO metric. This gives a reason for why detecting FMOs in X-ray images is challenging.

There are two main problems in deep learning-based real-time FMOs detection methods from high-resolution X-ray images, which are: 1) acceptable input image size of the current object detection framework is small or only square images can be used as the input. However, X-ray images of clothing packages for detecting small FMOs are in high resolution. Moreover, the processing time of current conveyor belt X-ray scanning solutions cannot meet real-time requirements and 2) small FMOs always appear blurred in the X-ray images due to a series of factors, such as the size of the optical center (focus of the X-ray emitter), the distance from the FMOs to the detector, the adjusted current and voltage, etc. Modern object detection methods such as Faster region-based convolutional neural networks (R-CNN), extract feature maps by applying max-pooling operations, which helps to downscale the input feature maps, to reduce the computational costs. However, the low spatial resolution causes difficulty in finding the precise location of the FMOs.

A well-designed FMOs detection method should tackle all of the above problems. This article expects to contribute to the online detection of small objects from X-ray images. The overall contributions of this work are summarized as follows.

1) We fused feature pyramid network (FPN) structure with the Faster R-CNN to better use the feature maps with higher resolution, which shows significant improvement on small FMOs detection.
2) We proposed a morphological erosion image down-sampling strategy to meet the real-time detection requirements and improve the performance degradation caused by the image down-sampling. Meanwhile, the integration of image preprocessing into the overall method facilitated real-time detection proves the promising practical application value of the proposed method in the engineering field.
3) We assessed the performance of framework with backbones of different depths. ResNet-50 is selected as an optimal backbone that achieves the speed and accurate balance for detecting FMOs from X-ray images. The trained model was tested for its robustness to the FMOs detection with a high recall rate.

II. LITERATURE REVIEW

A. Conventional X-Ray Object Detection Method

For decades, a significant amount of researches have been devoted to detecting objects from X-ray images by various image processing method. Hand-crafted feature extractors and

classifiers are widely used to detect targets from images. Zohora and Santosh [9], [10] proposed a circular foreign objects detection method based on edge detection algorithms and circular Hough transform (CHT). Morphological operations were applied to preprocess X-ray images. Their technique has achieved satisfactory detection accuracy in detecting button objects. However, noncircular foreign objects are ignored; meanwhile, at least 8.52 s are needed to process a single X-ray image, which may not be a feasible option for online product quality inspection. Santosh and Roy [11] proposed a sequential classifier-based approach for detecting arrows in biomedical images. A grating-based multimodal X-ray imaging method was used for novelty detection of foreign objects in food products [24]. Scale-invariant feature transform (SIFT) was used to detect illicit objects in X-ray images of luggage [25]. Although those researches have made remarkable achievements, traditional object detection method usually requires computer vision experts to craft effective features usually applicable for a particular scenario, thus proving the poor generalization ability.

B. Deep Learning-Based X-Ray Object Detection Method

Deep learning-based object detection architecture can be roughly grouped into two categories, which are two-stage detectors and one-stage detectors. Typical two-stage detectors, such as R-CNN [26], Fast R-CNN [27], Faster R-CNN [28], and spatial pyramid pooling net (SPP-Net) [29], generate predictions based on a region proposal outputted by selective search or CNN. The region of interest (ROI) [30] is captured at the first stage to output the target's potential locations. Then a classifier is applied only in the candidate regions. On the other hand, one-stage detectors, such as you only look once (YOLO) [31], RetinaNet [32], and single-shot detector (SSD) [33] are simple and straightforward as the detection runs directly on the original image without first identifying the ROI.

A notable amount of literature has been published on deep learning-based object detection in X-ray images. Santosh *et al.* [17] employed a two-stage detector, Faster R-CNN to detect circle-like foreign objects in chest X-rays and have achieved 97% precision, 90% recall, and 93% F1-score. Wei and Liu [6] proposed a multitask transfer learning approach based on SSD network to detect dangerous goods in X-ray images. The research focused on relearning the knowledge of detection models in object detection for visible light images. Liang *et al.* [34] examined the use of Faster R-CNN and SSD for automated detection of threats that can be found in airports. Jain [35] applied YOLOv2 and Faster R-CNN to perform threat object detection in X-ray images. Operiano *et al.* [36] proposed a genetic algorithm-based method to optimize YOLOv3 for specialized dangerous objects X-ray dataset, the formed networks have beaten YOLOv3 in accuracy without transfer learning. These researches have achieved outstanding success for object detection on various X-ray datasets. But these objects such as gun, Razor-blade, button, and knife are fairly large.

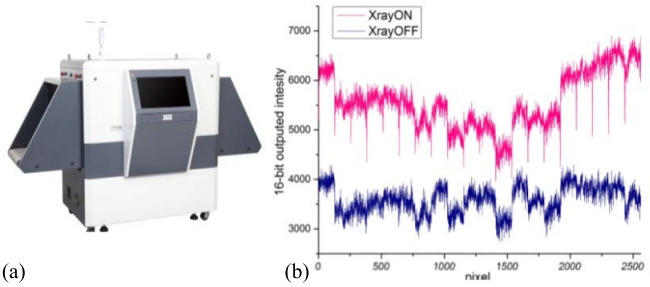

Fig. 2. X-ray-based FMOs detecting system and image normalization. (a) System appearance. (b) Offset values and fully illuminated values of photodiodes.

C. Small X-Ray Object Detection in High-Resolution Images

There are two different ways to tackle the problems in online detecting of small FMOs from high-resolution X-ray images based on deep learning. The first way is to run a sliding window classifier across the high-resolution X-ray image [37] to search for FMOs. Since multiple detection is usually required for each input X-ray image, overlaps between each window are involved to ensure detection of the boundaries, which can significantly increase processing time. These types of methods are almost unusable when applying a deep learning-based object detector, especially for real-time object detection with limited computing power in industrial occasions.

The second type of methods usually down-sample the high-resolution X-ray image to acceptable scales to the deep object detection model, so that each image can be detected only once. Although the image down sampling operations cause the loss of the small target details in the original image, the target enhancement preprocessing of the image before down sampling can improve the performance degradation problems. In this article, we propose a morphological erosion image down-sampling strategy to meet the real-time detection requirements and improve the detection accuracy. In YOLO, different from Faster R-CNN, only images with the same length and width are allowed to be input. Therefore, we combined Faster R-CNN with FPN to further the accuracy of small FMO detections. The trained model has been deployed to an industry computer for real-time experiment verifications, which can only provide limited computing power, and was equipped with GeForce GTX 750Ti.

This article is organized as follows. Sections III and IV introduced the proposed FMOs detection methodology. Section V discussed our experiments and empirical results. Section VI concluded the findings of this research.

III. Data Collection

A. X-Ray Imaging

The X-ray energy source works at 170 kV 0.8 mA, generated by an X-ray generator with a fan beam of 80° and RS232 standard control interface. A Sens-Tech XDAS-V3 system is applied for X-ray signal processing (SP) and data acquisition, which consists of 20 detector head (DH) boards to cover the 600 mm width conveyor belt and a SP board to sample the analog signal. Each DH board consists of

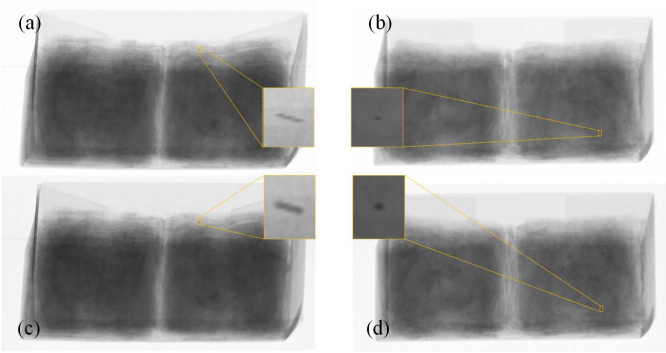

Fig. 3. Preprocessing comparison: (a) and (b) are the original X-ray images, while (c) and (d) are the corresponding images after preprocessing.

128-channel 0.4 mm pitch arrayed CsI scintillator photodiodes. Finally, the SP board samples the signal and provides line scans of $20 \times 128 = 2560$ pixels data via the integrated gigabit ethernet (GigE) interface. The data provided for each pixel are in 16-bit unsigned short format. The length of the package decides the width (columns) of the X-ray image. We selected an image width equal to 3200 in the experiment to cover packages with a maximum length of 1000 mm. It is worth noting there are blanks around the packages in the X-ray images.

Because of the differences in production and material, each photodiode of the DH board had different features. The data provided by the SP board need to be normalized by the following equation:

$$I = \text{abs}\left(\frac{I_{\text{raw}} - I_{\text{dark}}}{I_{\text{full}} - I_{\text{dark}}}\right) \times 255 \tag{1}$$

where I_{dark} is the offset value of photodiode without X-ray exposure. I_{full} is the value when photodiodes are under full illumination. I_{raw} is the value to be normalized, while I is the normalized value of a pixel in the range of [0, 255]. It is worth mentioning that I_{dark} and I_{full} were obtained at the beginning of the data sampling procedure. The X-ray-based FMOs detecting system and curves before normalization are shown in Fig. 2. The data accuracy was reduced from 16 to 8 bits to meet typical image format requirements in computer vision applications.

B. Dataset

In this research, 2000 gray scale 8-bit X-ray images with a resolution of 3200×2560 pixels were collected for experiments. Prepared FMOs were put in the packaging box with various positions and type combinations to ensure diversity. Since it is challenging to distinguish tiny FMOs from high-resolution X-ray images by human vision, each package was filled with five FMOs. All categories of FMOs were treated as one class to ensure the accuracy of annotation. Professionals from a clothing factory performed the annotations by using labeling software called LabelImg. A total of 1800 images with labels from 2000 were randomly selected to train the network. The remaining 200 were for testing. Dataset is available at https://github.com/chd-gr/FMOs_dataset.

C. Image Preprocessing

To preserve the features of FMOs in the X-ray images during the down-sampling [38], morphology erosion of a 5×5 ellipse kernel was applied to the raw X-ray images. Fig. 3 shows the original X-ray images and the images after the morphology preprocessing operation. Then the X-ray images were down-sampled from 3200×2560 to 1280×1024 with a constant aspect ratio.

IV. Modified Faster R-CNN

A. Backbone

The backbone has to be selected based on the types of objects to be detected, while the depth of the backbone depends on the complexity of the features to be extracted [39]. We selected ResNet-50 [40] as the backbone through experiments. This is a shallower model than ResNet-101 but much deeper than VGG-16 [41]. The ResNet-50 backbone, as shown in the top left of Fig. 4, is a deep stack of convolutional operations that consists of an input stem followed by four stages. Each stage consists of several bottlenecks (top right of Fig. 4) connected in series.

The backbone model computes a feature hierarchy stage by stage from the input images due to the max-pooling or convolutional operation [42]. Only feature maps with semantically strong features are used to detect the target objects. These methods have achieved good accuracy in detecting large-scale objects such as trees, cars, or buildings. However, to detect small FMOs in the X-ray images, the feature maps of stage 4 have been down-sampled, and their size is only 40×32 (see top right of Fig. 4). Therefore, the location accuracy of small FMOs in the feature maps is relatively low. By inputting an X-ray image with five FMOs in the left bottom of the trained ResNet-50 backbone, Fig. 5 shows some feature maps from stages 1–4 of different channels extracted by the backbone. It can be seen the feature maps of stage 1 have higher resolution but contain only low-level features, such as the edges. However, more precise locations of features can be provided by the feature maps of stages 2–3, which are critical for small object detection [43]. In the proposed improved Faster R-CNN, the FPN module combines feature maps with different resolutions that were outputted by stages 1–4 of the ResNet-50.

B. Feature Fusion by FPN

The FPN module is introduced to aggregate feature maps at different resolutions, as shown in the middle of Fig. 4. Formally, a list of multiscale feature maps $C = (C_1, C_2, C_3, C_4)$, all with 256 channels, where C_i were the feature maps extracted at stage i of the ResNet-50 then under a 1×1 convolution to unify the channels to 256. The FPN takes stages 1–4 input features and then they are added element-wise via lateral connections and up-sampling operations to generate fusion features as shown in the following equation:

$$\begin{cases} P_5 = \text{Maxpooling}(P_4) \\ P_4 = \text{Conv}_{3\times3}(M_4),\ M_4 = C_4 \\ P_3 = \text{Conv}_{3\times3}(M_3),\ M_3 = C_3 + \text{upsampling}(M_4) \\ P_2 = \text{Conv}_{3\times3}(M_2),\ M_2 = C_2 + \text{upsampling}(M_3) \\ P_1 = \text{Conv}_{3\times3}(M_1),\ M_1 = C_1 + \text{upsampling}(M_2). \end{cases} \tag{2}$$

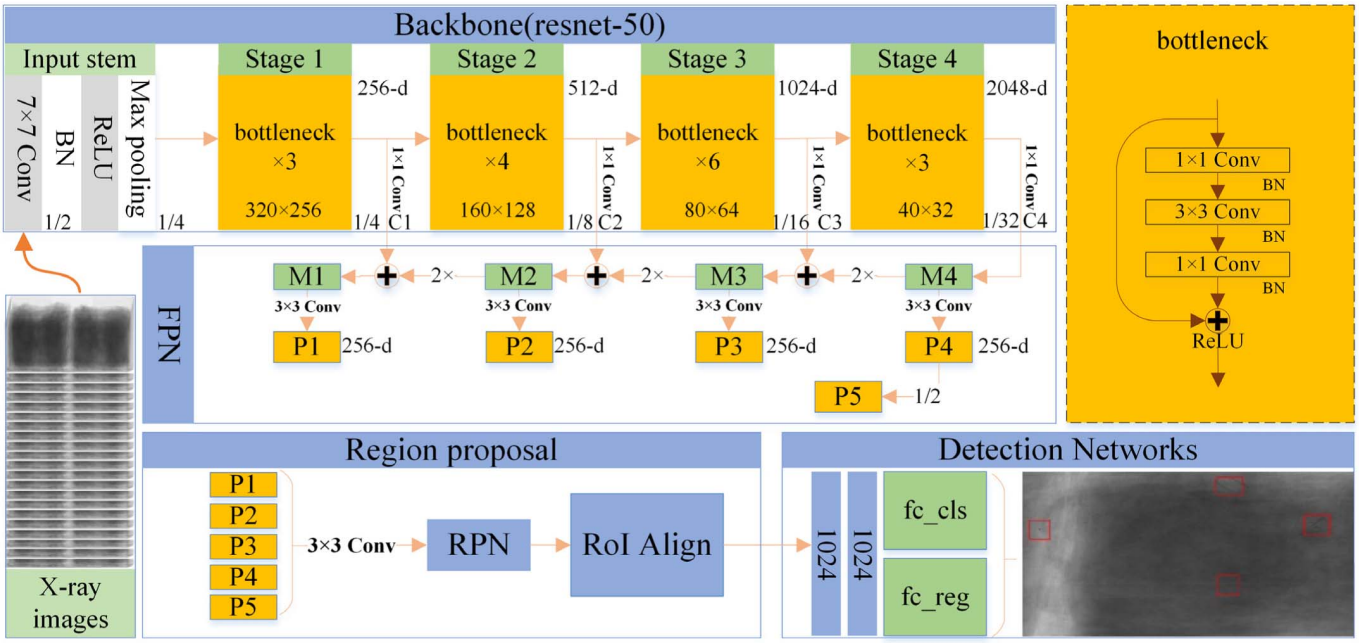

Fig. 4. Structure of the modified faster R-CNN framework.

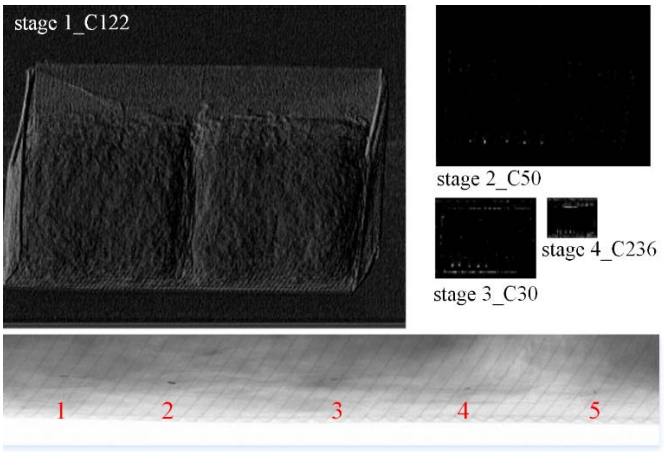

Fig. 5. Feature maps extracted from stages 1–4 of different channels by ResNet-50.

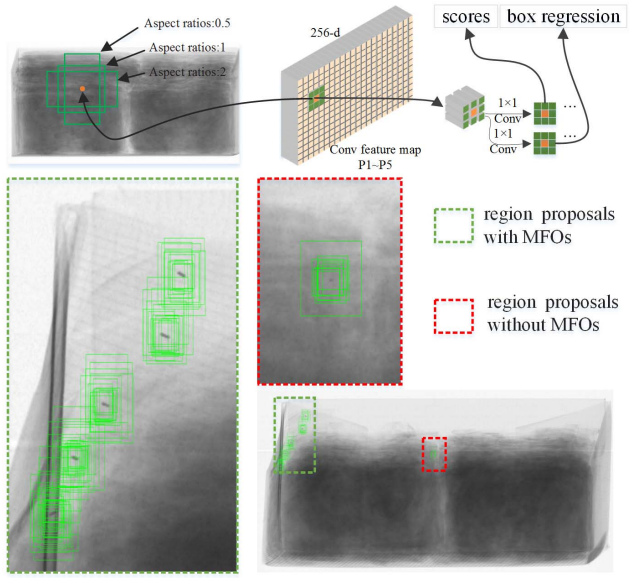

Fig. 6. Work flowchart of the RPN.

The lateral connections, which are similar to the U-net, also act as skip connections to make training easier [44].

C. Region Proposal Network

The region proposal network (RPN) is shown in the bottom left of Fig. 4. Feature maps produced by the FPN ($P1$–$P5$) are passed through a 2-D convolution layer, which contains 256 filters of size 3×3. The filters slide over each pixel of the feature maps, and the window is mapped to a lower dimensional feature. This feature is fed into two sibling 1×1 convolutional layers, which contain 3 and 12 filters to perform classification and regression tasks. Region proposals are created by anchors with aspect ratios of 0.5, 1, and 2 on the effective receptive field of the input image. Meanwhile, each pixel from feature maps $P1$–$P5$ can generate anchors with different sizes of 32, 64, 128, 256, and 512 according to the feature hierarchy. This process is displayed at the top of Fig. 6. The classification task assigns probabilistic labels to each raw proposal as positive or negative to decide whether the proposal contains the FMOs or not. For proposals that are deemed positive, the regression task fine-tunes the size to fit the dimensions of the FMOs. A set of rectangular object proposals were generated for the detector network. The bottom of Fig. 6 shows the generated proposals. It can be seen that proposals were generated around all five FMOs, shown as the area in the green box. False proposals generated around the interference object are shown in the red box.

Since massive coarse proposals will be generated by the RPN module, taking the feature map $P4$ as an example, the size of $P4$ (28×75) is 32 times smaller than the input image, $P4$ alone will generate $40 \times 32 \times 3 = 3840$ anchors. Many useless proposals will significantly increase the computation costs. If any given anchor has an intersection over union (IoU) greater than 0.7 with the ground truth (GT) box, we treat that anchor as a positive proposal. Similarly, if the IoU is smaller than 0.3, we treat the anchor as a negative region proposal. Also, if the IoU of a GT box between all candidate proposals is less than 0.7, the proposals with largest IoU will be selected as a positive proposal. Because many positive proposals would have a high overlap with one another. Nonmaximum suppression (NMS) [45] is employed to reduce the computation costs. The number of regression boxes were selected by NMS from 12 000 to 2000 proposals for further detection in the training stage. In the test stage, 256 proposals are taken from 2000 by the NMS.

D. Detector Network

The regions proposed by the RPN are fed into the ROI-Align layer [46] to convert the regions into a fixed size of 7×7. Then the fixed size ROIs are provided into two fully connected layers, each with 1024 neurons. The feature vectors are then inputted to another two fully connected layers to perform the classification and regression task. The classification assigns probabilistic positive and negative labels to each ROI, depending on the presence of the FMOs. The regression fine-tunes each positive-labeled box to match the FMOs dimensions and outputs the coordinates adjusted values of the proposed box. This process is displayed at the right bottom of Fig. 4.

E. Loss Function

The loss function is a critical item of a deep learning approach [47]. There are two primary objective functions for both the RPN and the detector network in the proposed modified Faster R-CNN framework. The first is a binary classification, and the other is a bounding box regression. We use the cross-entropy [48] loss to evaluate the binary classification

$$L_{\text{cls}}(y, \hat{y}) = -(y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y}) \tag{3}$$

where y is the GT label, while $\hat{y}$ is the predicted confidence of a bounding box containing FMOs. The second task loss L_{loc}, a smooth$_{L_1}$ loss function [49], is applied to quantify the location and size shift of the predicted bounding box $v = (v_x, v_y, v_w, v_h)$ between the GT box $t = (t_x, t_y, t_w, t_h)$

$$L_{\text{loc}}(t, v) = \sum_{i \in \{x, y, w, h\}} \text{smooth}_{L_1}(t_i - v_i) \tag{4}$$

where

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5, & \text{otherwise} \end{cases} \tag{5}$$

smooth$_{L_1}$ is a robust L_1 loss that is less sensitive to noises and outliers; meanwhile, L_2 loss is responsible for careful tuning of the minors shift to prevent exploding gradients. A multitask loss function L consists of the cross-entropy loss and the smooth$_{L_1}$ loss to jointly train for classification and bounding-box regression

$$L(y, \hat{y}, t, v) = L_{\text{cls}}(y, \hat{y}) + [y \geq 1] L_{\text{loc}}(t, v) \tag{6}$$

where the function $[y \geq 1]$ neglects all background ROIs while keeping the positive bounding boxes to reduce computation costs.

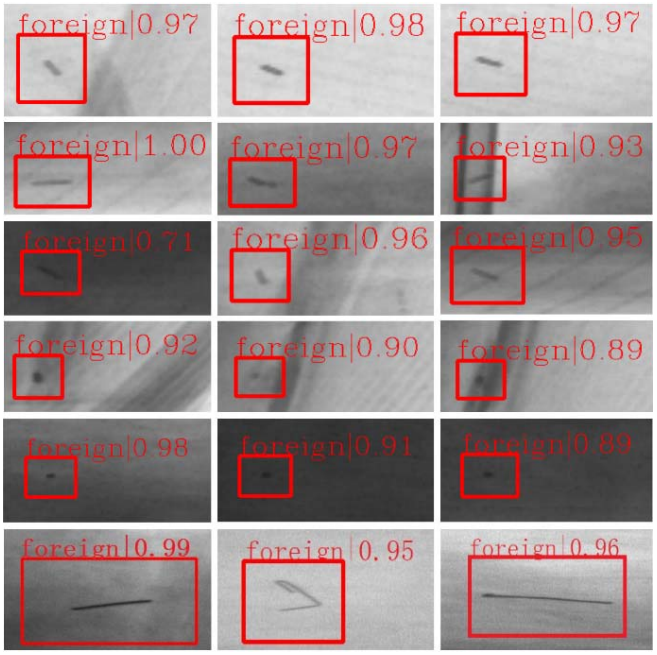

Fig. 7. Results of FMOs detected using the proposed approach: rows 1–3: broken needles; rows 4–5: steel balls with different sizes; and the last row: staple and wire.

V. Results and Discussion

The network was implemented in Python 3.6 programming language with the use of the Pytorch backend. The stochastic gradient descent (SGD) [50] optimizer (learning rate = 0.0025, momentum = 0.9, weight decay = 0.0001) was used to minimize the loss. The network was trained with 9000 steps, which took about 3.9 h on two GeForce RTX 2080Ti GPUs equipped with 32 GB of DDR5 RAM. Detecting a single test image by the trained network took 233 ms on average.

A. Experiment Results

Part of the visual results of the proposed approach is shown in Figs. 7 and 8. Some of the successfully detected FMOs, including broken needles, steel balls, staple, or wires in packaged clothing products, are displayed in Fig. 7. Meanwhile, some false positive (FP) and false negative (FN) cases are shown in Fig. 8.

By inspecting the visible results, it can be seen that four categories of FMOs with various positions were detected with high confidence. However, the texture and edges of the packages were of lower intensity, similar to the FMOs, which

leads to a few FP and FN detections. Two reasons cause this problem: first, the X-ray images' subsampling introduced the loss of details. For FMOs that are weak in appearance and close to the texture or edge of the package, their features are, therefore, submerged by interference. Second, the size of the GT box is larger than the size of the FMOs, which results in the GT box containing FMOs with much background, including some of the package edge and texture. This caused the model to consider some of the interference as characteristics of FMOs. However, the chance that FMOs that are weak in appearance are located just on the edge and texture of the packages is relatively small. Besides, this problem can be improved by reducing the size of the GT box or replacing the packaging box.

B. Performance of the Proposed Approach

The trained improved Faster R-CNN is tested by 200 X-ray images randomly selected from the dataset. Each image contained five FMOs, making a total of 1000 FMOs in the testing data. The trained model generates predicted bounding boxes on each test image. A threshold value for the IoU between the predicted and the GT bounding boxes decides whether the object detection is true or false. Object detection is challenging because the detection setting is highly adversarial. In general, it is challenging to ask a single classifier to perform uniformly well with all IoU levels. By inference, there is a tradeoff between recall and location.

In this research, finding more FMOs is our main aim. As a result, compared to the accuracy of location, a higher recall rate is desirable. The IoU metrics work well for large objects, but for small objects (e.g., 10×10 pixels) the IoU does not work well. This seems to be because small variations in the predicted pixel coordinates can make the overlaps significantly worse on small objects. The precision and recall curves at various conditions are shown in Fig. 9, the AP with IoU $= 0.75$ and 0.5 are 0.003 and 0.137, respectively. When the positioning error is removed by setting the IoU $= 0.1$, its AP grows to 0.946 as the blue area shown. Because the threshold of IoU is small now, it can be considered the positioning error is ignored, but it is guaranteed to find FMOs with not repeat detections. If all FPs are removed, then the AP grows to 0.950, and the remaining 0.05 is a FN. In general, the error of the trained Faster R-CNN with ResNet-50 + FPN mainly comes from the positioning error. In summary, the trained Faster R-CNN with ResNet-50 + FPN model detected 988 FMOs and missed 12. Meanwhile, one FP occurred. The model thus achieved a precision rate of 0.999, the recall rate is 0.988 and the F1-score is 0.993 on the testing dataset with IoU $= 0.1$.

C. Performance With Different Backbones

The Faster R-CNN with commonly used backbones of different depths was trained and tested on our dataset. The confusion matrix, precision, recall, F1-score, and testing foot per second results are listed in Table III. Compared with ResNet-50, ResNet-101 performs slightly better on small FMOs detection but has an increased calculation burden. However, the performance of the VGG16 model yields more FP and FN detections compared to the ResNet-50. We may draw the conclusion that the ResNet-50 backbone has enough accuracy in extracting the features of FMOs in the proposed X-ray images.

TABLE III
RESULTS WITH DIFFERENT BACKBONES

Framework		VGG16+FPN		ResNet-101+FPN		ResNet-50+FPN	
TP	FP	921	12	988	2	971	3
FN	TN	79	-	12	-	29	-
Precision		0.987		**0.998**		0.997	
Recall		0.921		**0.988**		0.971	
F1-score		0.953		**0.993**		0.984	
Testing fps		**4.8**		2.6		4.3	

TABLE IV
RESULTS WITH OR WITHOUT IMAGE PREPROCESSING AND FPN

Framework		ResNet-50+FPN		ResNet-50+FPN+Preprocessing		ResNet-50	
TP	FP	971	3	988	1	965	18
FN	TN	29	-	12	-	35	-
Precision		0.997		**0.999**		0.982	
Recall		0.971		**0.988**		0.965	
F1-score		0.984		**0.993**		0.973	
Testing fps		4.3		3.8		**5.0**	

D. Image Preprocessing and FPN

In this article, morphological image preprocessing, with the FPN, was applied to overcome the obstacles when detecting small FMOs by the commonly used deep learning methods. Table IV shows the Faster R-CNN with ResNet-50 + FPN and image preprocessing achieved the highest precision, recall rate, and F1-score. The recall rate improved from 0.965 to 0.971 with the FPN module. Simultaneously, the decrease in detection speed is relatively small, from 5.0 drop to 4.3 fps. On the other hand, image preprocessing further improved the model's recall and precision rate from 0.971 to 0.988 and 0.997 to 0.999, respectively. F1-score improved from 0.984 to 0.993.

E. State-of-the-Art Comparison

Further, the comparative with state-of-the-art methods has been made. Two well-known one-stage detectors, YOLOv4 and RetinaNet, were trained with the same training and testing dataset as our model. The size of the input images for RetinaNet and YOLOv4 is 1280×1024 and 608×608, respectively. The results are provided in Table V, where our method performs the best with precision, recall, and F1-score. It is worth mentioning that it took about three days to train the YOLOv4 network, but because of the limit on the input image size, the detection result was the worst.

F. Significance Test

McNemar's test [51] is used to analyse statistically whether the performance of the Faster R-CNN model is improved with the combination of the FPN module and the image preprocessing operation separately. To perform this test, the detection results of three models for each FMO are compared. The

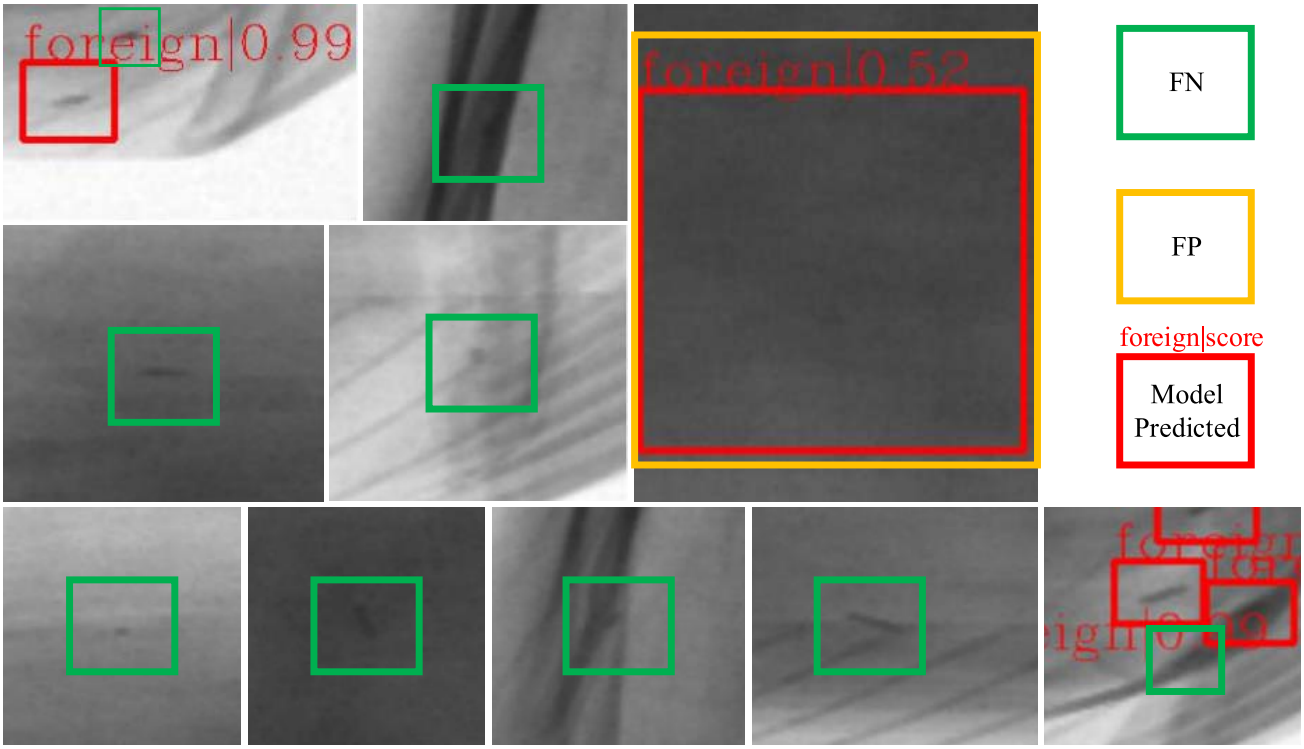

Fig. 8. Failure cases of FMOs detection. With nine FN (there exists an FMO, but the model fails to pick it out) and one FP (backgrounds had been detected as FMOs) cases displayed.

TABLE V
COMPARISON OF STATE-OF-THE-ART

Framework		RetinaNet		Our Method		YOLOv4@608×608	
TP	FP	963	11	988	1	893	21
FN	TN	37	-	12	-	107	-
Precision		0.989		**0.999**		0.977	
Recall		0.963		**0.988**		0.893	
F1-score		0.976		**0.993**		0.933	

contingency table is listed in Table V. Model A, Model B, and Model C are trained Faster R-CNN models with different frameworks; Model A: ResNet-50 + FPN + Preprocessing, Model B: ResNet-50 + FPN, Model C: ResNet-50. It can be seen intuitively from Table VI, the number of correct detections increased by 46 FMOs when added the FPN module, while image preprocessing further increased by 28 FMOs.

The two-tailed P values between model A and model B, model A and model C calculated with McNemar's test equals 0.0031 and 0.0001, respectively. The chi-squared equals 8.757 and 306.079 with one degrees of freedom. By conventional criteria, the differences are both considered to be statistically significant, which show the effectiveness of the proposed FPN module and X-ray image preprocessing method.

G. Online Testing

The trained model was deployed to an industrial control computer with i5-2500 CPU, 8 GB RAM, and GeForce GTX 750Ti for online testing. The online testing scenario and clothes are displayed in Fig. 10.

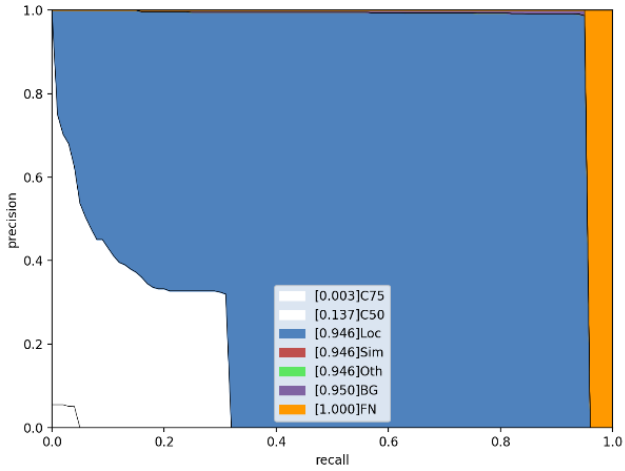

Fig. 9. Results with the COCO official analysis. C75: under the condition of IoU = 0.75, the area under the recall-AP curve; C50: under the condition of IoU = 0.5, the area under the recall-AP curve; Loc: under the condition of IoU = 0.1, background be misclassified as an object (BG): the area under the recall-AP curve after removing the FPs on the background.

For the four situations of broken needle, staple, steel ball, and no FMO inside, each situation was tested 384 times with clothes and FMOs filled to the package boxes randomly, 1536 times. Only one or none FMO is inside each test. The online testing generated eight FPs and five FNs, which mean 1147 of 1152 FMOs were recalled. A recall rate of 0.996 is obtained, which is even slightly higher than that on the offline test dataset. The results of online testing show the model has good generalization capability to clothes of different styles.

Fig. 10. Online testing scenario and the clothes of different styles used for online testing.

TABLE VI
CONTINGENCY TABLE

	Model B: Correct	Model B: Incorrect	Model C: Correct	Model C: Incorrect
Model A: Correct	963	28	962	46
Model A: Incorrect	9	4	4	7
Two-tailed P value	0.0031		0.0001	
Chi squared	8.757 with 1 degrees of freedom		33.620 with 1 degrees of freedom	

TP is treated as Correct while FP and FN are Incorrect.

VI. CONCLUSION

This article proposes an online detection approach, which is suitable for detecting small FMOs from X-ray images. A conveyor belt X-ray scanning system was developed for clothing package images collection. To overcome the obstacles of common deep learning-based detection methods for small object detection. FPN is combined with the Faster R-CNN to better use the feature maps with higher resolution, which shows significant improvement on small FMOs detection. The image down-sampling tactics is applied to meet real-time detection requirements. Morphological erosion is used to improve the performance degradation caused by the image down-sampling. Besides, backbones of different depths are evaluated; ResNet-50 is selected as an optimal one for detecting FMOs.

The proposed model achieved precision = 0.999, recall = 0.988, F1-score = 0.993, and AP = 0.946 on 200 offline test dataset. Experiments show the trained model can detect FMOs of different categories at various positions correctly. Especially for some FMOs with weak appearance, it is difficult for humans to distinguish even by zooming the image, yet the model can accurately identify them. The results of the proposed methods between the original Faster R-CNN are statistically significant by the McNemar's test. Meanwhile, the performance of the proposed method outperforms the state-of-the-art object detection method in precision and recall rate. Finally, online testing of the trained model reached a high recall rate of 0.996. In summary, the results show the effectiveness and robustness of the model for small FMOs detection from X-ray images, except for individual FMOs located on the edge of the packaging box with multiple layers of clothing stacked.

Although the proposed method can obtain satisfactory performance based on the collected dataset, some problems have not been addressed. First, the IoU of the predicted bounding boxes and the GT boxes is low. Second, the issue of including metal accessories in the clothing itself needs to be neglected. Therefore, more research should be conducted to improve the accuracy of small object positioning and eliminate interfering targets.

ACKNOWLEDGMENT

The authors would like to thank the editor and anonymous reviewers for their constructive comments, which helped to improve the quality of this article.

REFERENCES

[1] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: The rising trend of vision based measurement," *IEEE Instrum. Meas. Mag.*, vol. 17, no. 3, pp. 41–47, Jun. 2014.
[2] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *J. Big Data*, vol. 2, no. 1, pp. 1–21, Dec. 2015.
[3] V. Lashkia, "Defect detection in X-ray images using fuzzy reasoning," *Image Vis. Comput.*, vol. 19, no. 5, pp. 261–269, Apr. 2001.
[4] W. Du, H. Shen, J. Fu, G. Zhang, and Q. He, "Approaches for improvement of the X-ray image defect detection of automobile casting aluminum parts based on deep learning," *NDT E Int.*, vol. 107, Oct. 2019, Art. no. 102144.
[5] W. Du, H. Shen, J. Fu, G. Zhang, X. Shi, and Q. He, "Automated detection of defects with low semantic information in X-ray images based on deep learning," *J. Intell. Manuf.*, vol. 32, no. 1, pp. 141–156, 2021.
[6] Y. Wei and X. Liu, "Dangerous goods detection based on transfer learning in X-ray images," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 8711–8724, Jun. 2020.
[7] A. U. Fonseca, L. L. Oliveira, J. Mombach, D. S. A. Fernandes, R. Salvini, and F. Soares, "Foreign artifacts detection on pediatric chest X-ray," in *Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE)*, Aug. 2020, pp. 1–4.
[8] Z. Xue *et al.*, "Foreign object detection in chest X-rays," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2015, pp. 956–961.
[9] F. T. Zohora and K. Santosh, "Circular foreign object detection in chest X-ray images," in *Proc. Int. Conf. Recent Trends Image Process. Pattern Recognit.* Singapore: Springer, 2016, pp. 391–401.
[10] F. T. Zohora and K. C. Santosh, "Foreign circular element detection in chest X-rays for effective automated pulmonary abnormality screening," *Int. J. Comput. Vis. Image Process.*, vol. 7, no. 2, pp. 36–49, Apr. 2017.
[11] K. C. Santosh and P. P. Roy, "Arrow detection in biomedical images using sequential classifier," *Int. J. Mach. Learn. Cybern.*, vol. 9, no. 6, pp. 993–1006, Jun. 2018.
[12] F. T. Zohora, S. Antani, and K. Santosh, "Circle-like foreign element detection in chest X-rays using normalized cross-correlation and unsupervised clustering," *Proc. SPIE*, vol. 10574, Art. no. 105741V, Mar. 2018.
[13] B. Guan, J. Yao, G. Zhang, and X. Wang, "Thigh fracture detection using deep learning method based on new dilated convolutional feature pyramid network," *Pattern Recognit. Lett.*, vol. 125, pp. 521–526, Jul. 2019.
[14] L. Liu, M. Muelly, J. Deng, T. Pfister, and L.-J. Li, "Generative modeling for small-data object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6073–6081.
[15] I. D. Apostolopoulos and T. A. Mpesiana, "Covid-19: Automatic detection from X-ray images utilizing transfer learning with convolutional neural networks," *Phys. Eng. Sci. Med.*, vol. 43, no. 2, pp. 635–640, Jun. 2020.

[16] H. Iba and N. Noman, *Deep Neural Evolution: Deep Learning With Evolutionary Computation*. Berlin, Germany: Springer, 2020.
[17] K. C. Santosh, M. K. Dhar, R. Rajbhandari, and A. Neupane, "Deep neural network for foreign object detection in chest X-rays," in *Proc. IEEE 33rd Int. Symp. Comput.-Based Med. Syst. (CBMS)*, Jul. 2020, pp. 538–541.
[18] S. Yao, Y. Chen, X. Tian, and R. Jiang, "GeminiNet: Combine fully convolution network with structure of receptive fields for object detection," *IEEE Access*, vol. 8, pp. 60305–60313, 2020.
[19] S. Mascetti, D. Ahmetovic, A. Gerino, C. Bernareggi, M. Busso, and A. Rizzi, "Robust traffic lights detection on mobile devices for pedestrians with visual impairment," *Comput. Vis. Image Understand.*, vol. 148, pp. 123–135, Jul. 2016.
[20] N. Pradeep, H. Girisha, and K. Karibasappa, "Segmentation and feature extraction of tumors from digital mammograms," *Comput. Eng. Intell. Syst.*, vol. 3, no. 4, pp. 37–46, 2012.
[21] T. Zhao and R. Nevatia, "Car detection in low resolution aerial images," *Image Vis. Comput.*, vol. 21, no. 8, pp. 693–703, Aug. 2003.
[22] L. Steccanella, D. D. Bloisi, A. Castellini, and A. Farinelli, "Waterline and obstacle detection in images from low-cost autonomous boats for environmental monitoring," *Robot. Auton. Syst.*, vol. 124, Feb. 2020, Art. no. 103346.
[23] D. Chaudhuri, N. K. Kushwaha, A. Samal, and R. C. Agarwal, "Automatic building detection from high-resolution satellite images based on morphology and internal gray variance," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 5, pp. 1767–1779, May 2016.
[24] H. Einarsdóttir *et al.*, "Novelty detection of foreign objects in food using multi-modal X-ray imaging," *Food Control*, vol. 67, pp. 39–47, Sep. 2016.
[25] L. Schmidt-Hackenberg, M. R. Yousefi, and T. M. Breuel, "Visual cortex inspired features for object detection in X-ray images," in *Proc. 21st Int. Conf. Pattern Recognit.*, Nov. 2012, pp. 2573–2576.
[26] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
[27] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.
[28] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[29] P. Purkait, C. Zhao, and C. Zach, "SPP-Net: Deep absolute pose regression with synthetic views," 2017, *arXiv:1712.03452*. [Online]. Available: http://arxiv.org/abs/1712.03452
[30] R. N. Chityala, K. R. Hoffmann, D. R. Bednarek, and S. Rudin, "Region of interest (ROI) computed tomography," *Proc. SPIE*, vol. 5368, pp. 534–541, May 2004.
[31] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[32] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[33] W. Liu *et al.*, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.
[34] K. J. Liang *et al.*, "Toward automatic threat recognition for airport X-ray baggage screening with deep convolutional object detection," 2019, *arXiv:1912.06329*. [Online]. Available: http://arxiv.org/abs/1912.06329
[35] D. Jain and D. Kumar, "An evaluation of deep learning based object detection strategies for threat object detection in baggage security imagery," *Pattern Recognit. Lett.*, vol. 120, pp. 112–119, Apr. 2019.
[36] K. R. G. Operiano, H. Iba, and W. Pora, "Neuroevolution architecture backbone for X-ray object detection," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2020, pp. 2296–2303.
[37] A. Van Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," 2018, *arXiv:1805.09512*. [Online]. Available: http://arxiv.org/abs/1805.09512
[38] B. Cheng, Y. Wei, H. Shi, R. Feris, J. Xiong, and T. Huang, "Revisiting RCNN: On awakening the classification power of faster RCNN," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 453–468.
[39] K. S. V. Prasad, K. B. Dsouza, and V. K. Bhargava, "A down-scaled faster-RCNN framework for signal detection and time-frequency localization in wideband RF systems," *IEEE Transactions on Wireless Communications*, vol. 19, no. 7, pp. 4847–4862, Jul. 2020.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[42] J. Huang *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7310–7311.
[43] L. Zhu *et al.*, "Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 121–136.
[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[45] J. Hosang, R. Benenson, and B. Schiele, "Learning non-maximum suppression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4507–4515.
[46] M. Teichmann, M. Weber, M. Zollner, R. Cipolla, and R. Urtasun, "MultiNet: Real-time joint semantic reasoning for autonomous driving," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 1013–1020.
[47] J. Huyan, W. Li, S. Tighe, Z. Xu, and J. Zhai, "CrackU-Net: A novel deep convolutional neural network for pixelwise pavement crack detection," *Struct. Control Health Monitor.*, vol. 27, no. 8, p. e2551, Aug. 2020.
[48] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.
[49] K. Senhaji, H. Ramchoun, and M. Ettaouil, "Training feedforward neural network via multiobjective optimization model using non-smooth L1/2 regularization," *Neurocomputing*, vol. 410, pp. 1–11, Oct. 2020.
[50] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, Y. Lechevallier and G. Saporta, Eds. Berlin, Germany: Physica-Verlag, 2010, doi: 10.1007/978-3-7908-2604-3_16.
[51] M. Eliasziw and A. Donner, "Application of the McNemar test to non-independent matched pair data," *Statist. Med.*, vol. 10, no. 12, pp. 1981–1991, Dec. 1991.

Rong Gao received the B.Sc. and M.Sc. degrees from Chang'an University, Xi'an, China, in 2010 and 2014, respectively, where he is currently pursuing the Ph.D. degree with the School of Information Engineering.

His current research fields include machine vision; artificial intelligent, and deep learning techniques for traffic information engineering and control.

Zhaoyun Sun received the M.S. degree in computer application from Xi'an Highway University, Xi'an, China, in 1991, and the Ph.D. degree in road and railway engineering from Chang' an University, Xi'an, in 2007.

She has been a Professor with the School of Information Engineering, Chang'an University, since 2003. Her research interests include intelligent traffic condition detection and information processing, digital image processing, and traffic information engineering and control.

Ju Huyan received the B.Sc. and M.Sc. degrees from Chang'an University, Xi'an, China, in 2013 and 2016, respectively. She is currently pursuing the Ph.D. degree with the Centre for Pavement and Transportation Technology (CPATT), Waterloo, ON, Canada.

She joined the School of Transportation, Southeast University, Nanjing, China, in 2021. Her current research fields include pavement management; pavement condition evaluation; 3-D (2-D) image processing pavement distress detection; artificial intelligent (AI), machine learning, and deep learning techniques for pavement condition assessment and evaluation.

Wei Li received the B.S. and M.S. degrees in opto-electronic technique from Xidian University, Xi'an, China, in 2003 and 2006, respectively, and the Ph.D. degree in physical electronics from the University of Chinese Academy of Sciences, Xi'an, in 2009.

His research interests include digital image processing, automatic pavement condition detection and evaluation, machine learning, deep learning, and so on.

Liyang Xiao was born in Shaanxi, China, in 1996. She received the B.S. degree from Chang'an University, Xi'an, China, in 2018, where she is currently pursuing the master's degree with the School of Information Engineering.

Her current research fields include artificial intelligence, deep learning, and image processing.

Bobin Yao (Member, IEEE) received the B.Sc. degree in communication engineering and the M.Sc. degree in traffic information engineering and control from Chang'an University, Xi'an, China, in 2005 and 2008, respectively, and the Ph.D. degree from the Institute of Information Engineering, Xi'an Jiaotong University, Xi'an, in 2014.

He joined as a Faculty with the School of Electronic and Control Engineering, Chang'an University, in July 2014. His general research interests lie in the areas of statistical signal processing and convex optimization, including radar, communications and navigation, multi-in–multi-out (MIMO) wireless communication system, target localization and tracking, wireless networks, and intelligent transportation system (ITS).

Huifeng Wang was born in 1976. He received the Ph.D. degree in engineering from Xidian University, Xi'an, China, in 2009.

He is a Professor with Chang'an University, Xi'an, where he is the Director of Provincial Electrical and Electronic Experimental Demonstration Center. His research interests include intelligent transportation system (ITS), traffic environment perception, photoelectric detection and image processing.

Dr. Wang is a member of the special committee of Internet of Things Technology, Shaanxi Electronic Society and a member of the China Association of Automation and China association of optical engineering. He is a Reviewer of several magazines including IEEE ACCESS, *Optics and Laser Technology*, IEEE SENSORS, *Sensor Review*, *Transactions of the Institute of Measurement and Control*, and so on.